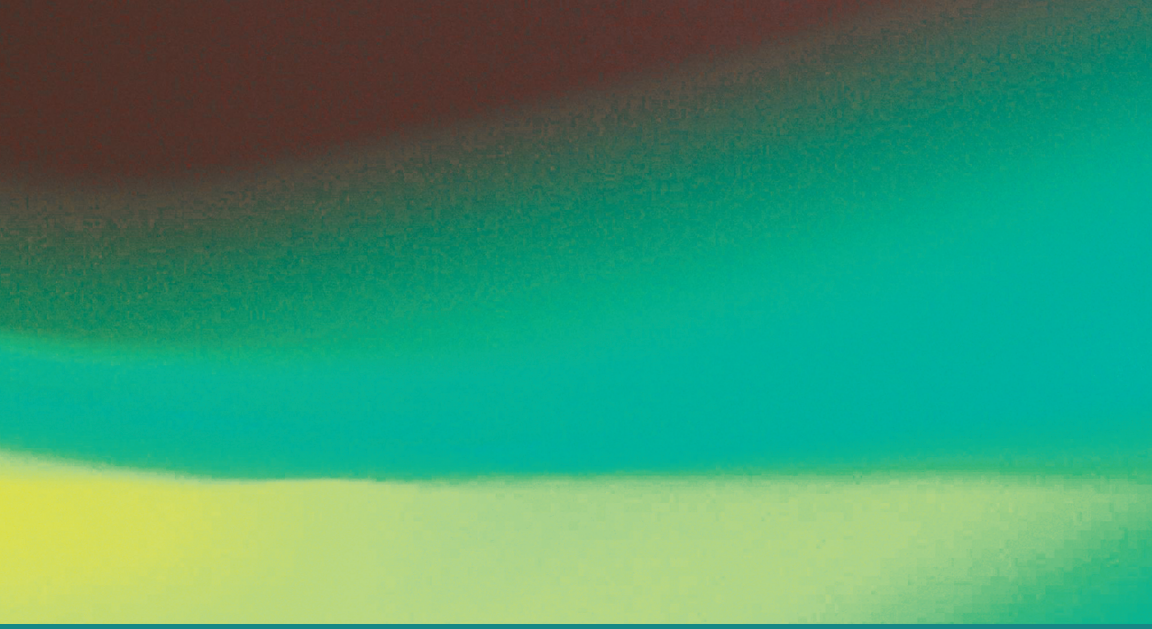

*Routledge Studies in Contemporary Philosophy*

# MENTAL ACTION AND THE CONSCIOUS MIND

Edited by Michael Brent and Lisa Miracchi Titus

# Mental Action and the Conscious Mind

Mental action deserves a place among foundational topics in action theory and philosophy of mind. Recent accounts of human agency tend to overlook the role of conscious mental action in our daily lives, while contemporary accounts of the conscious mind often ignore the role of mental action and agency in shaping consciousness. This collection aims to establish the centrality of mental action for discussions of agency and mind. The thirteen original essays provide a wide-ranging vision of the various and nuanced philosophical issues at stake. Among the questions explored by the contributors are:

- Which aspects of our conscious mental lives are agential?
- Can mental action be reduced to and explained in terms of non-agential mental states, processes, or events?
- Must mental action be included among the ontological categories required for understanding and explaining the conscious mind more generally?
- Does mental action have implications for related topics, such as attention, self-knowledge, self-control, or the mind-body problem?

By investigating the nature, scope, and explanation of mental action, the essays presented here aim to demonstrate the significance of conscious mental action for discussions of agency and mind. *Mental Action and the Conscious Mind* will be of interest to scholars and graduate students working in philosophy of mind, philosophy of action, and philosophy of agency, as well as to philosophically inclined cognitive scientists.

**Michael Brent** is a Responsible AI Expert at Boston Consulting Group (BCG). Prior to emerging from the cloistered walls of the academy, he was an Assistant Professor of Philosophy at the University of Denver, USA. His published work has appeared in the *Canadian Journal of Philosophy, Journal of Buddhist Ethics*, and *Philosophical Psychology*.

**Lisa Miracchi Titus** is an Associate Professor of Philosophy at the University of Pennsylvania, USA, where she is also a General Robotics, Automation, Sensing, and Perception (GRASP) Lab Faculty Affiliate and a MindCORE Faculty Affiliate. Her published work has appeared in the *Journal of Philosophy, Journal of Artificial Intelligence and Consciousness*, *Philosophical Psychology*, *Philosophy and Phenomenological Research*, and *Synthese*.

## Routledge Studies in Contemporary Philosophy

**Perspectives on Taste**
Aesthetics, Language, Metaphysics, and Experimental Philosophy
*Edited by Jeremy Wyatt, Julia Zakkou, and Dan Zeman*

**A Referential Theory of Truth and Falsity**
*Ilhan Inan*

**Existentialism and the Desirability of Immortality**
*Adam Buben*

**Recognition and the Human Life-Form: Beyond Identity and Difference**
*Heikki Ikäheimo*

**Autonomy, Enactivism, and Mental Disorder**
A Philosophical Account
*Michelle Maiese*

**The Philosophy of Fanaticism**
Epistemic, Affective, and Political Dimensions
*Edited by Leo Townsend, Ruth Rebecca Tietjen, Hans Bernhard Schmid, and Michael Staudigl*

**Mental Action and the Conscious Mind**
*Edited by Michael Brent and Lisa Miracchi Titus*

**Epistemic Injustice and the Philosophy of Recognition**
*Edited by Paul Giladi and Nicola McMillan*

For more information about this series, please visit: https://www.routledge.com/Routledge-Studies-in-Contemporary-Philosophy/book-series/SE0720

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# Mental Action and the Conscious Mind

Edited by Michael Brent and
Lisa Miracchi Titus

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2023
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

© 2023 selection and editorial matter, Michael Brent and Lisa Miracchi Titus; individual chapters, the contributors

The right of Michael Brent and Lisa Miracchi Titus to be identified as the author[/s] of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

*Trademark notice:* Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

*Library of Congress Cataloging-in-Publication Data*
A catalog record for this title has been requested

ISBN: 978-0-367-07751-8 (hbk)
ISBN: 978-1-032-07115-2 (pbk)
ISBN: 978-0-429-02257-9 (ebk)

DOI: 10.4324/9780429022579

Typeset in Sabon
by KnowledgeWorks Global Ltd.

**For Louisa, whose mental actions continue to astonish and delight –M.B.**

**For Noah and Neva, who continually teach me how rewarding jumping in midstream can be. –L.T.**

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# Contents

*List of contributors* xi
*Preface* xii

**Introduction** 1
LISA MIRACCHI TITUS AND MICHAEL BRENT

1 **Disappearing Agents, Mental Action, Rational Glue** 14
JOSHUA SHEPHERD

2 **How to Think Several Thoughts at Once: Content Plurality in Mental Action** 31
ANTONIA PEACOCKE

3 **Attending as Mental Action: Mental Action as Attending** 61
WAYNE WU

4 **The Most General Mental Act** 79
YAIR LEVY

5 **Mental Action and the Power of Effort** 100
MICHAEL BRENT

6 **Inference as a Mental Act** 122
DAVID HUNTER

7 **Reasoning and Mental Action** 142
MARKOS VALARIS

8 **Causal Modeling and the Efficacy of Action** 164
HOLLY ANDERSEN

**9 Skepticism about Self-Understanding** 182
MATTHEW BOYLE

**10 Embodied Cognition and the Causal Roles of the Mental** 205
LISA MIRACCHI TITUS

**11 Two-Way Powers as Derivative Powers** 228
ANDREI A. BUCKAREFF

**12 Are Practical Decisions Mental Actions?** 255
ALFRED R. MELE

**13 Self-control, Attention, and How to Live without Special Motivational Powers** 272
SEBASTIAN WATZL

*Index* 301

# Contributors

**Holly Andersen,** Simon Fraser University, Canada

**Matthew Boyle,** University of Chicago, U.S.A.

**Michael Brent,** Boston Consulting Group, USA

**Andrei Buckareff,** Marist College, U.S.A.

**David Hunter,** Ryerson University, Canada

**Yair Levy,** Tel Aviv University, Israel

**Alfred Mele,** Florida State University, U.S.A.

**Antonia Peacocke,** Stanford University, U.S.A.

**Joshua Shepherd,** Carleton University, Canada

**Lisa Miracchi Titus,** Lisa Miracchi Titus, University of Pennsylvania, U.S.A.

**Markos Valaris,** University of New South Wales, Australia

**Sebastian Watzl,** University of Oslo, Norway

**Wayne Wu,** Carnegie Mellon University, U.S.A.

# Preface

This project was originally conceived and undertaken by Michael Brent. Lisa Miracchi Titus was brought on to help bring the project to fruition. For this reason, some previous references to work in this volume as forthcoming refer to Michael Brent as the sole editor. Please list both editors in future citations. Part of the work performed by Lisa Miracchi Titus was supported by a Fellowship from the National Endowment for the Humanities.

We wish to thank the contributors for their patience, and for the hard and careful work of our graduate assistants John Roman, Maja Sidzinska, and Jacqueline Wallis.

# Introduction

*Lisa Miracchi Titus and Michael Brent*

We make decisions, remember, reason, pay attention. These are things we mentally do, mental actions. Mental actions pervade our lives, but the very idea of a mental action raises many questions. What makes an action *mental*, per se? Is there any important difference between the kinds of actions that involve overt behavior and the ones that don't? And what makes mental actions *actions*? What's the difference between mental processes that aren't actions and ones that are? How, exactly, does the agent figure in as a cause of action and how could this be compatible with a physicalist and empirically informed worldview? And then, for any kind of mental action, such as remembering or reasoning, specific questions arise about the ways a subject's agency is involved, and what its limitations are. For example, what is required for inference to be a mental action? Does it need to be intentional? What relation does the agent bear to the inference, and to the resultant belief? Are decisions about what to do actions? Is attention somehow fundamental to mental action? If so, how?

The ubiquity of mental action, and its connection to consciousness and agency, would suggest that the topic deserves a foundational place in action theory and philosophy of mind. However, it has received less attention than it deserves.[1] Although it is beyond the scope of this introduction to fully explain why this is the case, we think that different emphases in philosophy of mind and action have contributed to the neglect of mental action per se. It is our intention in this volume to bring together a variety of voices that together will show how multi-faceted, important, and relevant the topic of mental action is to current debates. Moreover, we think that the time is right to revisit the topic of mental action: developments in philosophy and relevant sciences over the last several years motivate a fresh look at the topic of mental action.

In what follows, we first lay out the terrain, providing some reasons why there has been a tendency to overlook mental action in philosophical literature of the past several decades and raise some concerns for this omission. We then describe some of the features of recent philosophical and scientific developments that may bear on questions concerning

DOI: 10.4324/9780429022579-1

the nature of mental actions, their characteristic features, explanations appealing to them, and scientific explanations of them. Then we provide an overview of the main themes of the collection, and describe how the different contributions bear on these themes. We conclude by articulating some lessons to be learned from the collection as a whole which provide direction for future research.

## 0.1 The State of Play for Mental Action

One reason why mental action has fallen through the cracks over recent decades has to do with emphases in discussions of consciousness and action that have the effect of leaving mental action out. Discussions about *phenomenal consciousness* have tended to focus on sensory experience, rather than agential experience, as the paradigm case. Discussions of *mental processes* have tended to focus on mental processes underlying them in ways that do not appeal to the agent as such. Discussions of *action* have tended to focus on action that manifests in behavior. As these are three of the most relevant philosophical topics to the topic of mental action, the result has been not only an omission of the topic of mental action, but sometimes even genuine puzzlement about how mental action might fit into mental and agential taxonomies. Of course, any brief discussion that generalizes about broad literature cannot do justice to their complexity or nuance. However, it is instructive to look at the "forest" for a moment, to see how each of these literatures might have framed their questions in ways that exclude mental actions. In doing so we can better understand how the articles in this collection contribute to the re-integration of mental actions in discussions in philosophy of mind and action and can help these literatures develop and expand as well.

Contemporary accounts of the conscious mind often focus on experience, highlighting the sensory aspects of consciousness and what it is like to undergo certain conscious experiences. In such discussions, the usual examples of conscious mental states that possess phenomenal properties are often assumed to be non-agential, such as perceptual experiences and bodily sensations.[2] Moreover, attempts to reductively explain consciousness often aim to identify it or its grounds with some kind of representational or computational property (such as information integration, or existence in a global workspace), which is specifiable independently of any reference to agency.[3] Intentionally or not (no pun intended), a predominant focus on sensory as opposed to agential manifestations of consciousness, along with an emphasis on reductive characterizations of consciousness, have downplayed the existence of mental action as a paradigm case of conscious experience.

Similarly, philosophical work on the nature of mental processes has tended to focus on questions of reduction, and so has aimed to specify

mental processes in neutral functional, computational, or other more fundamental terms.[4] A key aspect of these reductive projects has been to eliminate any essential reference to an agent, replacing talk of agency with talk of the "cognitive system". Attention, reasoning, and decision-making, then, might be understood independently from any appeal to an *agent* who attends, reasons, or makes decisions. The category of mental action is then uninteresting, if one is interested in the nature of mental processes. Agency is always inessential to their most perspicuous characterization.

Perhaps as the other side of the coin on this issue, action theory tends to overlook the topic as well, focusing instead on actions that involve overt behavior.[5] This may also be due to the dominant reductionist accounts of intentional action, on which actions are behaviors that are appropriately caused by intentions. As such, the very idea of a mental action may seem ill-placed, or at least of a very different category than intentional behavior.

Interesting recent work motivates a re-evaluation of these emphases on sensation, reduction, and behavior. In particular, philosophical work in philosophy of mind and action has taken a non-reductive turn, where mental kinds are understood to inherently involve the environment and normative evaluation. For example, the knowledge-first turn motivates a view of mental kinds as inherently world-involving, and so not reducible to internal states or processes.[6]

Additionally, on this view, the world-involvingness of behavioral actions does not set them apart from mental actions. *Both* are inherently world-involving. Moreover, there is a growing appreciation of the importance of agency for mental kinds that are widely agreed to be paradigmatic of both mentality and consciousness. Attention, for example, is increasingly understood to be an inherently agential phenomenon and also to be bound up with perception and particularly perceptual consciousness.[7] Furthermore, increasingly theorists are examining metaphysical questions about the nature of agency, and its relation to causation, with fresh eyes. There may be ways to make sense of how agents can be ineliminably implicated in mental actions that are not ontologically costly, or do not challenge physicalist commitments.

This collection aims to establish the centrality of mental action for discussions of agency and mind. We have brought together thirteen original essays by both leading scholars in the field and up-and-coming philosophers to show how interesting and nuanced the various philosophical issues are, and how various advances in both philosophy of mind and action theory motivate re-examining the topic of mental action and investigating it directly. We also hope to show that focusing on mental action directly can inform our understanding of many other philosophical issues, including consciousness, intentionality, agency, the ontology of mental kinds, and the nature of causation.

## 0.2 Themes of the book

Rather than structuring the book with papers corresponding to discrete themes, the papers in this collection each touch on a variety of themes in overlapping and complementary ways, motivating the interest and depth of the topic. We have thus decided to introduce the items in this collection by describing their relevance to these themes.

### *0.2.1 What is mental action?*

All papers in this volume touch on the question of the nature of mental action, and several specifically address it. Some provide a conception of mental action that illuminates it as playing a special role in our mental lives. For example, **Joshua Shepherd**, in "Disappearing Agents, Mental Action, Rational Glue", argues that mental actions such as deliberation, decisions, and reflection upon reasons have the functional role of navigating internal practical conflict. In having this role, they help us to understand the way the agent is distinctively involved as such in her mental life. It is this role that ties together mental actions and the applicability of normative assessment, such as responsibility and rationality.

**Antonia Peacocke** identifies a striking and philosophically important feature of some intentional mental actions: they have distinct contents under their distinct intentional descriptions, i.e., they have what she calls *content plurality*. She uses the example of adding up the cost of your meal, which consists of a $25 appetizer and a $34 entrée. She argues that when you use your understanding that the total cost of the meal is the sum of these two numbers, you're adding $25 and $34 *just is* your figuring out the cost of the meal. Your mental action has distinct contents—one content as an adding-up, and another content as a figuring-out. In particular, it has the propositional content *25 + 34 = 59* as an act of addition, and the content *$59* is the cost of the meal as an act of figuring out the cost of the meal. She argues that there is nothing further that one must do in order to judge that $59 is the cost of the meal. She then uses content plurality to argue that some mental actions are *both* decisions and judgments, contrary to the standard approach to taking these to be exclusive categories.

**Wayne Wu**, in "Attending as Mental Action. Mental Action as Attending", argues that attention is the cognitive mechanism by which the agent uses one among many potential targets of attention to guide her behavior. Wu conceptualizes actions as certain kinds of trajectories through the space of possible behaviors for the agent. Trajectories that are actions involve the agent's control over the trajectory, through (minimally), selection of the guiding input. Mental actions can be conceptualized similarly, where the output is not a bodily movement, but either a modification of the initial mental state or a new mental state.

**Yair Levy,** in "The Most General Mental Act" comes at the issue of attention and mental action from a different angle. Taking inspiration from Williamson's claim that knowledge is the most general factive mental state, Levy argues that attention is the most general mental act – it is a determinable of determinate mental actions. This explains its ubiquity and centrality to action while accounting well for the heterogeneity of its different manifestations.

Others resist standard attempts to reductively explain mental processes, and in particular mental actions. For example, **Michael Brent**'s "Mental Action and the Power of Effort" argues that mental actions crucially involve the agent as cause, and that the way this occurs involves the exertion of effort by the agent towards her end. He is clear that this is intended to be a non-reductive account of mental action.

Yet others leverage advances in our understanding of causation and metaphysical ontologies in order to illuminate the nature of mental causation. For example, **David Hunter**, in "Inference as a Mental Act", defends the claim of his chapter's title, and in so doing also motivates a conception of the agent as the sustaining and motivating cause of her action. He puts pressure on the centrality of willed action to thinking about action, broadening the traditional conception of agential powers. According to Hunter, a mental act such as inferring is an action because it has an agent as its cause, not because it is backed by some further feature of agency, such as willing or decision. In a similar vein, **Markos Valaris**, in "Reasoning and Mental Action", challenges the view that inferring is a process. Inferring is widely considered to be a mental action, and Valaris argues that we can vindicate that claim without understanding inferring to be a process. In particular, it is the control we have over our inferred beliefs in virtue of their being held for reasons that makes inference a mental action.

### *0.2.2 In what ways do explanations involving mental actions go beyond typical causal explanations?*

One central issue for understanding mental actions is the issue of the relationship between explanations that appeal to mental actions and causal explanations that do not inherently involve appealing to personhood or agency.

**Holly Andersen**, in "Causal Modeling and the Efficacy of Action", argues that rationalizing action-explanations have a kind of normativity that is not exhausted by causal explanations. She compares work on mathematical explanations to partially explain what is not captured by causal explanation. She argues, however, that in some cases we can treat action-explanations as though they were causal explanations, and in so doing make reliable predictions. For example, in a case where one is kneading bread and then letting it rise in order to make a loaf of bread,

the kneading does not cause the rising. Still, against the background of the end of making bread, kneading predicts rising, and so the correlations between them can be analyzed using tools for causal modeling.

**Matthew Boyle**, in "Skepticism about Self-Understanding", also weighs in on this issue, arguing that the kinds of reason-explanations that we regularly give when we take ourselves to transparently know our reasons for an attitude or action are not processualist, resists what he calls the "ballistic" conception of mental action, where are the reasons that we cite can be understood merely as prior causes that can be disconnected from our current worldview. He distinguishes between causal explanations that are "extrinsic" in the sense that we are not citing causes of what causally determined our current total mental state, but instead, we are citing features of that very same current total state that causally sustain other features, and thus are "intrinsic". It is a feature of rational agency, Boyle argues, that we can regularly give these intrinsic explanations. **Lisa Titus** (née Miracchi) argues that the inherent meaningfulness of our mental actions and other cognitive processes entails that they are inherently embodied and environmentally embedded. They therefore cannot be identified with or reduced to intracranial neural or computational processes, though they may ultimately be explainable in terms of a coordination of intracranial processes, body, and environment.

**Andrei Buckareff** offers a framework to help us understand how an agent settles whether or not to make a decision when in a context of practical uncertainty. For Buckareff, understanding how this occurs requires understanding the sorts of things that agents are. He believes that agents are not enduring substances, but functionally integrated systems composed of many simple objects and their causal powers. Understood in this way, when an agent settles whether or not to make a particular decision, it is the activity of the total agglomeration of simple objects and their causal powers that produces the decision. **Joshua Shepherd** argues that the agent enters into action-explanation not by being an agent cause, or by being the agent of a special kind of motivational state or event, but through the abilities to reason and plan that organize such a complex system. As such, looking for the special involvement of the agent at the token action-causation level is misguided. The participation of the agent is more holistic and temporally extended.

### *0.2.3 Can mental action be reduced to or explained in terms of non-agential mental states or events?*

In different ways, several authors in the collection grapple with the question of whether mental actions can be reduced to non-agential mental states or events, or that a research program aiming to do this will illuminate the nature of mental action. In addition to the contributions from Brent and Titus above, **Holly Andersen** highlights the normativity of

action in general and the ways in which attempts to treat action explanations as causal explanations can obscure special normative features of action explanations. Moreover, in her defense of the usefulness of causal modeling tools for some action explanations, she makes salient the ways in which certain structuring features that are not included in the model (such as the final end of baking bread in the example just described) are required for the applicability of causal modeling tools. As such, she helps to provide a nuanced picture of the ways in which tools for explaining non-mental, non-agential phenomena can be applicable to mental action without motivating reductive theses.

**Andrei Buckareff** takes the opposing view, focusing on a class of non-reductionist views about intentional action according to which intentional agency involves the exercise of irreducible two-way powers: for example, powers to make it true that p or to make it true that not-p. Buckareff challenges those (e.g., Alvarez, 2013; Mayr, 2011; Steward, 2012) who believe that when an agent settles whether or not to make a particular decision, they do so by exercising a unique two-way power, viz., the power either to perform an action or not. Against such views, Buckareff introduces two objections. First, he argues that they have difficulty explaining *why* a two-way power is manifested the way it is in light of an agent's reasons for deciding in a particular way. Second, he argues that commitment to the existence of such two-way powers requires commitment to substance dualism. He then offers a reductive account that draws on Molnar's (2003) account of "derivative" powers to show how two-way powers might be derived from one-way powers.

### *0.2.4 Is attention fundamental to mental action?*

Attention is a central and ubiquitous type of mental act, one whose nature has attracted considerable interest from both psychologists and philosophers for over a century. In his chapter, **Sebastian Watzl** shows that attention has a central role in all mental agency. His argument is focused on self-control. He provides an account of self-control that challenges a dominant trend across recent work in psychology and philosophy. That trend explains self-control in terms of special motivational powers, such as will-power or a division between a deliberative and an emotional motivational system. Watzl argues that no such special motivational powers are necessary. Instead, Watzl argues, self-control illustrates the importance of the mental activity of attention in the control of all action. Attention is important for self-control, Watzl argues, because attention is an agential capacity by which agents can actively couple or decouple an intention, preference, or desire to and from action. Attention, Watzl argues, acts as a flexible interface between our standing states and our actions. Self-control is achieved through a complex set of attentional skills. He calls this the re-prioritization account of self-control. Self-control, according

to him, thus indicates a behavioral flexibility or freedom that attention weaves into the structure of all agency.

**Yair Levy** sets out and defends a novel account of attention, according to which *attending* is the most general type of mental act, that which one performs on some object if one performs any mental act on it at all. In his view, attending is entailed by every mental act one that performs. Levy describes the entailment of attention by mental action as a datum that any credible theory of attention should accommodate and explain. After clarifying and defending this core datum, Levy offers an explanation for why this thesis holds, viz., each particular mental act-type one performs is said to be a determinate of the determinable attention. This is the sense in which attention is the most general mental act: if one is attending to some object O, then one is performing some more determinate mental act on O. Along the way to developing this account of attention, Levy criticizes two prominent yet different explanations of attention, offered by Christoper Mole and Wayne Wu, respectively.

**Wayne Wu** theorizes about both intentional action and attention using the concept of a behavior space and trajectories through it. According to Wu, we can understand intentional action as the constraining of an agent's path through the behavior space by her intention. Attention is the agent (her mind) taking possession of an input to the behavior space in order to constrain the path through that space. In this way, attention is illuminated as being in the service of, and partially constitutive of, intentional actions. This view can be extended to consider mental actions as well. Wu considers certain kinds of mental illnesses such as schizophrenia and depression in terms of an incapacity to constrain behavior space. Mental actions, then, for Wu, just are trajectories through behavior space whose output is non-bodily (non-muscular). (Covert) attention can then be understood both as constitutive of the actions it participates in and as a mental action itself.

### *0.2.5 To what extent does science challenge whether there's a phenomenon of genuine mental action?*

**Matthew Boyle** argues against empirically-motivated skepticism about self-knowledge and self-understanding. The body of empirical research in question shows that people sometimes confabulate their reasons for making decisions or performing actions. For example, Nisbett and Wilson (1977) show that subjects will tend to confabulate in experimental conditions where they are told to choose between (what are in fact identical) stockings. There is a strong position effect, such that subjects are more likely to choose a stocking on the right. However, subjects are unaware of this bias, and tend to cite features of the chosen stocking as the reasons for their decision. (Boyle also discusses evidence from split-brain patients.) This kind of evidence is often adduced as support for

the view that we do not have the kind of transparent knowledge of our reasons that we often take ourselves to.

Boyle defends this commonsense conception of our own agency by distinguishing between processualist and sustaining-cause explanations. A processualist explanation cites a cause in the past as a determinant of a subsequent attitude or action. A sustaining-cause explanation, in contrast, cites a concurrent sustaining cause of the attitude or action. So, for example, an explanation of why one came to like one's friend may be different from the reasons *now* for which one likes them. Boyle claims that the empirical evidence only casts doubt on our ability to provide processualist explanations but not on sustaining-cause explanations. The person who chose the right-hand stocking really does, at the moment they are asked, take it to have a better sheen or color (etc.) than the others. This latter kind of explanation is what we have transparent access to (not infallibly, but as a rule), as a feature of our nature as rational agents.

**Al Mele**, in "Are Practical Decisions Mental Actions?", also resists the skeptical philosophical conclusions based on empirical experiments. He offers a variety of interpretations of Libet-style experiments, showing how nuanced the issues of both phenomenology and interpretation are. He argues for the existence of genuine practical decisions that resolve uncertainty about what to do, and argues that their existence is not undermined by the Libet experiments. Moreover, he adduces additional empirical considerations from Schurger, Sitt, and Dehaene (2012), who offer a stochastic accumulation model that predicts the Libet results. According to Schurger et al. (2012), the neural readiness potential Libet detected and that many have interpreted as a decision to move one's finger is in fact just stochastic "build up" which is highly correlated with intentionally moving one's finger but has not yet crossed the threshold of committing to action.

**Wayne Wu** implicitly argues for the exact opposite, using science as argues that theorizing about the psychological and biological reality of attention, as opposed to conceptual analysis, helps us develop a more apt account of the phenomenon, and so of mental action altogether. Moreover, in contrasting his view with the account of perceptual attention as a kind of "spotlight" on perceptual information, he shows how important methodological and other framework commitments are in even framing the space of plausible hypotheses.

**Lisa Titus** also appeals to empirical considerations in order to refine our philosophical views of mental processes. She argues that taking seriously the ineliminability of semantic vocabulary in the higher-level sciences (psychology, social science, animal behavior, economics) supports the view that semantic contents are causally relevant to mental processes, thus requiring us to revise intracranial reductionist accounts. She also provides one plausible evolutionary account of how systems might have evolved to have this *semantic efficacy*.

In all four cases, the authors motivate a nuanced and global approach to empirically informed philosophy, and try to make clear methodological assumptions that are often implicit in research practice. In doing so they put pressure on a common tendency to treat robust accounts of mental action in conflict with scientific evidence.

### *0.2.6 What implications does mental action have for related issues, such as agential responsibility, conscious experience, and self-knowledge?*

**David Hunter** defends the view that agents are causally responsible for what they mentally do, and that this can help to explain why agents are responsible for what they believe, and can be creditworthy for their beliefs. He distinguishes his account of causal responsibility from more widespread appeals to intention in action, arguing that the broader category still is important for ethical concerns. (We can be held responsible for forgetting to do something we promised to do, for example.) **Markos Valaris** weighs in on this issue by arguing that we can explain the responsibility an agent has for what she believes without conceiving of the agent, or her reasons, as temporally prior causes of her inferred beliefs.

In defending his view of practical decisions as the intentional resolving of uncertainty, **Al Mele** explores questions of phenomenology. Is the experience of actionally acquiring an intention distinguishable from the experience of non-actionally acquiring it? He defends the view that there is an important phenomenal difference. Roughly, this corresponds to the difference between "picking" one of many options versus just taking one. Actionally acquiring an intention to A is also phenomenally distinct from acquiring a conscious urge to A, for one could have the urge to A and still pick a different option.

In addition to Matthew Boyle's contribution, which bears on self-knowledge as described above (§2.5), **Antonia Peacocke** argues that the content plurality of mental actions offers a new explanation of self-knowledge. She argues that a single mental action can both be that *p* and a judgment that *I believe that p*. Thus some mental actions involve knowledge of themselves. She argues that the same is true for knowledge of what one intends. The same mental action can both be a decision to A and a judgment that one intends to A.

## 0.3 Future directions

In investigating the nature and scope of mental action, the essays presented here demonstrate the depth and interest of the topic, and motivate including mental action among the central topics in philosophy of mind and action. As this brief overview hoped to show, studying mental action is fascinating

and worthwhile in its own right, and it has implications for a wide range of issues from the scope and import of causal and scientific explanations, to our understanding of phenomenology and the grounds of agential responsibility. The variety of positions included here, as well as the number of interconnections between the different contributions, demonstrates just how rich and ready for exploration the field of mental action now is.

We close this introduction by drawing together a few lessons from the collection that point in promising directions for future research. First is the importance of investigating mental action directly, without immediately assimilating it to or subsuming it under other mental, agential, or more fundamental processes. Mental actions have important intentional, phenomenological, and agential characteristics that deserve further illumination and articulation (Andersen, Brent, Mele, Titus, Peacocke). Secondly, issues concerning ontological categories are not straightforward, and deserve careful attention. The kind of responsibility that we attribute to mental actions may not require a process-based account of action. Instead, we may have to look more holistically at the whole agent (Boyle, Buckareff, Hunter, Valaris). Attempts to view mental action as a kind of process should not be taken for granted. Third, there are interesting opportunities for new reductive approaches, for example, those that take into account the structure of the whole agent (Shepherd, Wu) or build upon recent work in metaphysics (Watzl). We look forward to seeing the development of work on mental action beyond this volume in these and hopefully many more fruitful directions.

## Notes

1. O'Brien and Soteriou (2009) is an edited collection of essays on the topic. Book-length treatments are scarce. See Soteriou (2013), Proust (2014), and Watzl (2017).
2. See the extensive literature inspired by Nagel (1974), Jackson (1982), and Chalmers (1996).
3. See e.g. See Tye (2000), Tononi (2004), and Baars (2005).
4. For some contemporary accounts see Carruthers (2015) and Shea (2018).
5. Three recent anthologies together contain merely two papers that directly examine mental action. For example, see Aguilar, Buckareff, and Frankish (2010) and Dancy and Sandis (2015). The former has twelve papers of which a single entry addresses the topic, while the latter contains thirty-seven papers none of which directly address mental action. In addition, of the seventy-five entries in O'Connor and Sandis (2010), only one discusses the notion of mental acts. And there are but three book-length discussions of the topic: Soteriou (2013), Proust (2014), and Watzl (2017). O'Brien and Soteriou (2009) is an edited collection of essays on the topic.
6. See Williamson (2000), Nagel (2013), Miracchi (2015), Kelp (2017), Ichikawa and Jenkins (2017), Simion (2018) for relevant developments of the knowledge-first approach. For the application of knowledge-first ideas to action theory, see Levy (2013, 2016), Miracchi and Carter (forthcoming).
7. See Burge (2010), Carrasco (2011), Prinz (2012).

## References


Aguilar, J. H., Buckareff, A. A., & Frankish, K. (Eds.) (2010). *New waves in philosophy of action*. Palgrave-Macmillan.

Alvarez, M. (2013). Agency and two-way powers. *Proceedings of the Aristotelian Society*, *113*, 101–121.

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience? *Progress in Brain Research*, *150*, 45–53.

Burge, T. (2010). *Origins of objectivity*. New York: Oxford University Press.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press.

Dancy, J., & Sandis, C. (2015). *Philosophy of action: An anthology*. West Sussex, UK: Wiley-Blackwell.

Ichikawa, J., & Jenkins, C. S. I. (2017). On putting knowledge 'first'. In J. A. Carter, E. C. Gordon, & B. W. Jarvis (Eds.), *Knowledge first: Approaches in epistemology and mind*. Oxford: Oxford University Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, *32*(127), 136.

Kelp, C. (2017). Knowledge first virtue epistemology. In J. A. Carter, E. C. Gordon, & B. W. Jarvis (Eds.), *Knowledge first: Approaches in epistemology and mind*. Oxford: Oxford University Press.

Levy, Y. (2013). Intentional action first. *Australasian Journal of Philosophy*, *91*(4), 705–718.

Levy, Y. (2016). Action unified. *The Philosophical Quarterly*, *66*(262), 65–83.

Mayr, E. (2011). *Understanding human agency*. New York: Oxford University Press.

Miracchi, L. (2015). Competence to know. *Philosophical Studies*, *172*(1), 29–56.

Miracchi, L., & Carter, J. A. (forthcoming). Refitting the mirrors: On structural analogies in epistemology and action theory. *Synthese*.

Molnar, G. (2003). *Powers: A study in metaphysics*. New York: Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–50.

Nagel, J. (2013). Knowledge as a mental state. *Oxford Studies in Epistemology*, *4*, 275–310.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. *Psychological Review*, *84*(3), 231–259.

O'Brien, L., & Soteriou, M. (Eds.) (2009). *Mental actions*. Oxford: Oxford University Press.

O'Connor, T., & Sandis, C. (Eds.) (2010). *A companion to the philosophy of action*. West Sussex, UK: Wiley-Blackwell.

Prinz, J. (2012). *The conscious brain: How attention engenders experience*. New York: Oxford University Press.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. New York: Oxford University Press. DOI:10.1093/acprof:oso/9780199602162.001.0001

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, *109*(42), E2904–E2913.

Shea, N. (2018). *Representation in cognitive science*. New York: Oxford University Press.

Simion, M. (2019). Knowledge-first functionalism. *Philosophical Issues*, *29*(1), 254–267.

Soteriou, M. (2013). *The mind's construction: The ontology of mind and mental action*. New York: Oxford University Press.

Steward, H. (2012). *A metaphysics for freedom*. New York: Oxford University Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(42).

Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.

Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. New York: Oxford University Press.

Williamson, T. (2000). *Knowledge and its limits*. New York: Oxford University Press.

# 1 Disappearing Agents, Mental Action, Rational Glue[1,2]

*Joshua Shepherd*

## 1.1 How agents disappear

The problem of the disappearing agent arises, for some, in the context of reflection upon event-causalist accounts of intentional action (Brent, 2017; Velleman, 1992; Wallace, 1999).[3] According to such accounts (Brand, 1984; Mele, 1992), intentional action is behavior caused in the right way by the acquisition of certain mental states (e.g., intentions). The worry is that talk of events and states within an agent cannot amount to action by the agent.

### *1.1.1 Velleman*

David Velleman complains that the 'standard story' of intentional action leaves something out: "my objection is that the occurrences [the standard causalist view] mentions in the agent are no more than occurrences in him, because their involvement in an action does not add up to the agent's being involved" (1992, p. 463).

The standard causalist has an obvious response, and Velleman considers it. It is that the standard story contains the agent implicitly:

> The reasons, intention, and movements mentioned in the story are modifications of the agent, and so their causal relations necessarily pass through him. Complaining that the agent takes no part in causal relations posited between reasons and intention, they might claim, is like complaining that the ocean takes no part in causal relations posited between adjacent waves.
>
> (Velleman, 1992, p. 463)

Velleman rejects this response, offering two cases.[4] He first references Frankfurt's unwilling addict. This is a person who has a first-order desire in favor of taking a substance, and a second-order desire in favor of dropping the first-order desire. The addict takes the substance because of the first-order desire's strength. The addict takes it unwillingly because

DOI: 10.4324/9780429022579-2

the action runs counter to the second-order desire. Velleman comments that "being the subject of causally related attitudes and movements does not amount to participation of the sort appropriate to an agent" (1992, p. 463). Velleman's view is that an account of action that includes only mental states and causal relationships between them makes the agent disappear—there is a thing called an agent, it participates in action, this participation only takes place when standards are met, and these standards go beyond the normal (non-deviant) causal operations of normal mental states (e.g., intentions and beliefs).

Velleman further illustrates this view with a different kind of case. In this case, a character meets with an old friend "for the purpose of resolving some minor difference" (1992, p. 464). Things have been complicated, such that during this meeting subconscious intentions to sever the friendship 'crystallize,' and the meeting ends with both characters angry at each other. Velleman intends this as a case in which action occurs – there is non-deviant causation of behavior by appropriate mental states. Even so, in this case, the agent disappears. Velleman claims: "Surely, I can believe that the decision, though genuinely motivated by my desires, was thereby induced in me but not formed by me; and I can believe that it was genuinely executed in my behavior but executed, again, without my help" (1992, pp. 464–465).

This case is different in important ways from that of the unwilling addict. What they have in common is that both cases capitalize on *agentive complexity* and the *internal practical conflict* that may result. The addict has desires at cross-purposes. The angry friend has intentions at cross-purposes—the explicit intention to meet in order to resolve a difference, and the subconscious intention to end the friendship.

Velleman's proposal is sensitive to the need to resolve internal practical conflict.[5] He argues that we need a mental state to play the functional role appropriate to the agent's participation in action. This role is that of reflecting on available reasons for action and taking sides with the best reasons. Velleman then argues that the agent is involved in the action when these mental actions of reflection and taking sides are motivated by a special kind of mental state—a desire to act in accordance with reasons.

> We say that the agent turns his thoughts to the various motives that give him reason to act; but in fact, the agent's thoughts are turned in this direction by the desire to act in accordance with reasons. We say that the agent calculates the relative strengths of the reasons before him; but in fact, these calculations are driven by his desire to act in accordance with reasons. We say that the agent throws his weight behind the motives that provide the strongest reasons; but what is thrown behind those motives, in fact, is the additional motivating force of the desire to act in accordance with reasons. For when a

> desire appears to provide the strongest reason for acting, then the desire to act in accordance with reasons becomes a motive to act on that desire, and the desire's motivational influence is consequently reinforced. The agent is moved to his action, not only by his original motive for it, but also by his desire to act on that original motive, because of its superior rational force. This latter contribution to the agent's behaviour is the contribution of an attitude that performs the functions definitive of agency; it is therefore, functionally speaking, the agent's contribution to the causal order.
>
> (Velleman, 1992, p. 479)

Very few have accepted Velleman's proposal. Later I discuss what I think has gone wrong. But there is something interesting and insightful here. When Velleman reaches for the role of the agent, he reaches for certain kinds of mental actions—deliberation, decision, reflection upon reasons broadly construed (which is plausibly constituted by mental actions of many types—counterfactual thinking, prospective imagination, attempts to remember alternatives, attempts to infer likely consequences, etc.). These are mental actions that have a function of navigating internal practical conflict.

### *1.1.2 Wallace*

Let us bookmark this point, and turn to R. Jay Wallace's (1999) discussion of the disappearing agent. Like Velleman, he is concerned with cases of internal practical conflict, and these cases stem from agentive complexity of the sort humans exemplify. For Wallace is motivated by cases of akrasia, and in particular by cases of action motivated by addiction.

Wallace contrasts two models of the will. He calls the first the hydraulic model. This model depicts desires "as vectors of force to which persons are subject, where the force of such desires in turn determines causally the actions the persons perform" (Wallace, 1999, p. 630). The concept of desire in play excludes intentions—one can desire something without intending or choosing it. But on the hydraulic model, the agent always does what she most desires to do. In this way, desire plays the key role in causal explanations of action.

> They determine which action we perform by causing the bodily movements that we make in acting, the assumption being that the strength of a given desire is a matter of its causal force in comparison to the other given desires to which we are subject.
>
> (Wallace, 1999, p. 631)

Wallace's primary complaint about this model is that it "leaves no real room for genuine deliberative agency" (1999, p. 633). Cases of akrasia

motivate this complaint. For on the hydraulic model, there is no room to maintain that the agent could have done otherwise than act against her best judgment, unless the strength of her desires were different. Wallace comments: "*Given* the causal strength of the various desires to which they are actually subject, together with their actual beliefs, it turns out that *akratic* agents simply lack the capacity to do what they judge best" (1999, p. 633).

So, Wallace complains that a picture of agency that does not permit agents to rationally handle cases of internal practical conflict—a desire goes one way, while a judgment goes another, and the rational move is to follow the judgment—makes the agent disappear. Citing Velleman, Wallace says the following: "Action is traced to the operation of forces within us, with respect to which we as agents are ultimately passive, and in a picture of this kind real agency seems to drop out of view" (1999, p. 633).

Recall that at this stage, Velleman posited a novel kind of desire to stand in for the agent by motivating a range of rationality-promoting mental actions. Wallace's proposal is slightly harder to parse. It involves a different motivational state, and it appears to involve, as well, a novel capacity. According to Wallace's *volitionalist model*, the special motivational state is a volition—"a kind of motivating state that … [is] directly under the control of the agent" (1999, p. 636). Wallace offers intentions, decisions, and choices as sub-types of volitions. So, these special motivational states are defined in terms of a special capacity—the agent's direct control over them. It is in these states and this capacity, according to Wallace, that we find agency itself:

> [I]ntentions, decisions, and choices are things we do, primitive examples of the phenomenon of agency itself. It is one thing to find that one wants some chocolate cake very much, or that its odor reminds one of one's childhood in Detroit, quite another to resolve to a piece. The difference, I would suggest, marks a line of fundamental importance, the line between the passive and the active in psychological lives.
>
> (1999, p. 637)

One should worry that appeal to an agent's direct control over a special class of motivational states obscures more than it illuminates. How are we to understand the capacity for direct control? Is it always rational? If so, then it is redundant—judgments about what it is best to do would seem to be enough. If not—and since it is a capacity humans possess, this is the option we have to take—we lack an explanation for why it operates rationally at times, and irrationally at others.

But detailed criticism of Wallace's proposal is not my aim here. It is the pattern of reasoning that interests me.

### *1.1.3 Brent*

Like Velleman and Wallace, Michael Brent looks at event-causalist accounts of intentional action and sees disappearing agents. He is not motivated, however, by cases involving agentive complexity or internal practical conflict. It seems that no special rational state or capacity classified in terms of event causation will satisfy him. Brent argues that we need a new view of causation. According to Brent, it is not states or events that act, it is agents, and no event-causal account could "properly account for your causal role when you are initiating, sustaining, and controlling the movements of your body during an action" (2017, p. 665). What is required to capture this is "a plausible alternative conception of causation" (Brent, 2017, p. 668).

But this does not seem to be all. Brent endorses non-reductive agent causation. But he also endorses the posit of a special causal power that amounts to agent causation. This causal power manifests in the exertion of effort.

> I suggest that the fundamental difference between the bodily movements that you are making happen during an action and those movements that are merely happening is that the former are occurring in conjunction with your exertion of effort, whereas the latter are not. Although the bodily movements you are making happen during an action and those that are merely happening might seem to be instances of the same type, they are not. The causal contribution of your exertion of effort differentiates those movements that you are making happen while acting from those that are merely happening, marking the movements as categorically distinct.
>
> (Brent, 2017, p. 667)

As with Velleman and Wallace, one might worry that this proposal raises more problems than it solves. We lack reasons to think that effort uniquely produces intentional actions, as opposed to a wide range of effects. Nothing about a notion of substance causation in general could help here, since even if substance causation is fundamental, lots of non-agentive substances cause lots of non-actional effects. And there seems nothing especially illuminating in the notion of effort itself. It seems plausible that effort could sometimes cause non-actional byproducts. (The event-causalist might suggest that we introduce intentions here, to distinguish between effects that amount to intentional actions and effects that do not. Brent would reject this move.) We also have reason to think that a problem of deviant causation afflicts this proposal as much as it does event-causalist accounts. For the conjunction of effort and bodily movements have to be stitched together somehow.

Again, though, it is not detailed criticism that is my aim, but the pattern of reasoning at work.

### *1.1.4 The moral of the story so far*

In my view, Velleman's, Wallace's, and Brent's proposals are responsive to a genuine problem, namely, how to understand the relationship between the agent and the actions the agent produces (or executes). Further, both Velleman's and Wallace's proposals are sensitive to a deep truth about the nature of agency. The deep truth is the close association of agency and the application of rational standards to behavior.

Like Velleman, Wallace is motivated by cases of internal practical conflict. Such cases, if handled improperly, seem to make the agent disappear. Similar to Velleman, Wallace posits a special class of mental state that could bring the agent back into the picture. In addition, he commits to a special mental capacity of direct control over these states. Both theorists are, then, guided by an image of the true core of agency.

Brent does not acknowledge the importance of practical norms in understanding the nature of agency. But he is viewing a version of this image as well. For Brent, the true core of agency involves a causal power to produce effects by exerting effort.

All three theorists are, in different ways, committed to capturing the core of agency in terms of some very special item – a state, capacity, or mode of causation. And all three execute this commitment at the level of the production of individual actions. They seem to want to shoehorn the agent into the action at the level of action explanation. This leads them to posit esoteric items that could play a role in action explanation.

All three theorists, and the many who have found the problem of the disappearing agent compelling, are misled. An alternative picture of agents is required.

## 1.2 The nature of agency

The problem of action is, quoting Frankfurt, "to explicate the contrast between what an agent does and what merely happens to him" (1998, p. 69). As I understand it, it is a contrast between kinds of event: between *action* on the one hand, and *mere behavior* on the other. I am focused here on a different—though importantly related—problem. We might call it the problem of agency, and understand the challenge as one of explicating the contrast between what qualifies as an agent and what does not. As I understand it, it is a contrast between kinds of system: between *agents* on the one hand, and *non-agential systems* on the other.

A number of philosophers have found some attractive versions of the idea that to qualify as an agent a system should conform to certain normative standards. Often this is put in terms of rationality. So, for example, Donald Davidson claims that "An agent cannot fail to comport most of the time with the basic norms of rationality" (2004, pp. 196–197). And Christian List and Phillip Pettit claim that "The very idea of an

agent is associated with some standards of performance or functioning, which we call 'standards of rationality.' These must be satisfied at some minimal level if a system is to count as an agent at all" (2011, p. 24).

Some argue for a more encompassing notion of agency. It will be useful to look briefly at it.

Consider a very simple system, only capable of moving in one direction along a flat surface. It has some internal structure. It has some causal powers. We might impute a function to this system, based on its survival needs: the system needs to find and fall into small gaps in a surface on which it moves. If it fits into the gap, it wins. (Say it avoids predators, or finds food.) The system does so in the only way it can, by moving blindly in one direction along the wall. There aren't so many small gaps in the wall, but every once in a long while—make it as unlikely as you like – it comes across one. It wins. This may be enough for the system. Such a system does not trend in the direction of agency.

Now consider a slightly more complex system—the paramecium, a single-celled eukaryote. It has some internal structure. It has some causal powers. It moves through certain liquids by the coordinated beating of its cilia. It can navigate around obstacles or escape certain substances by way of an "avoiding reaction"—a reversal of its direction of movement, and a slight changing of course in some random direction (Kung & Saimi, 1982). This is not a very efficient way of navigating, but the thing is stuck with very short cilia. In any case, it is also capable of reproducing, and its methods appear good enough for evolution to keep it employed—many a paramecium has survived long enough to reproduce.

Some think that unlike our earlier system, the paramecium trends in the direction of agency. Tyler Burge argues that "primitive agency" extends down to the level, at least, of single-celled eukaryotes. Burge points to the orientation behavior of such organisms:

> Taxes are directional movements with respect to stimulations in the environment. They require sensory capacities that are directional. Usually determining direction depends on there being two or more locations of sensory receptors on the body of the organism. Directional movement is usually achieved by some mechanism in the animal for simultaneous differentiation of intensities of stimulus registration in different bodily sensors. For example, the animal might turn toward or away from the direction perpendicular to the side of its body that receives the most intense stimulus registration.
>
> (2009, p. 258)

Burge judges that coordinated, functioning orientation behavior of simple organisms—e.g., "The paramecium's swimming through the beating of its cilia, in a coordinated way, and perhaps its initial reversal

of direction" (2009, p. 259)—qualify them as agents. As Burge writes, "Such organisms are capable of steering toward or away from a stimulus source, subsequent to internal differentiations between stimulus intensities in different areas of the body" (2009, p. 258). The movement toward a stimulus is caused in a different way than the movement away from a stimulus, and the difference makes sense in light of the system's own activity—the transitions between states of the system that are differentially sensitive to stimulus source and intensity. That is, the system's behavior is not only reliably produced, but also coherently produced given the circumstances. And it permits something like success. The system's behavior is related to imputable goals regarding its needs (for safety, for finding energy sources, or whatever) with respect to its environment. In its typical behavioral circumstances, this orientation behavior reliably leads to successful (enough) approximation of these goals.

Many will disagree with Burge that we find agency at this level. After all, reproduction is no less complicated and important a process for the paramecium than is locomotion. But it is less intuitive to think of a paramecium's asexual reproduction, by a process of binary fission, as an example of primitive agency. That's just the mechanics of life. And if so, why not the beating of the cilia, or the avoiding reaction (which, by the way, often occurs spontaneously)?

Whatever we think about the agency of a paramecium, Burge is right to emphasize continuity between this level and others. At this low level, we find key ingredients of agency. First, behavioral standards—standards of success—must be imputable to the system. This trends in the direction of the application of rational norms. Second, behavior must be coherent in light of the relevant behavioral standards. This trends in the direction of rational behavior. Third, behavior must be reliable in meeting or approximating these standards—the system must succeed, to some degree. This trends in the direction of control.

Look at systems more internally sophisticated (and usually more causally powerful) than the paramecium. Behavioral standards that apply to the system can still be drawn from the system's needs or functions. But at this level, the system begins to set certain standards for itself. For at this level the system has the capacity to represent objects, and goals for action regarding these objects—to token psychological states that link it and its behavior to the world in reliable ways. Burge here invokes the notion of a perspective: "When perception sets an object for animal action, agency reaches a new level of sophistication. The action is suited to a goal that the animal itself perceptually represents. If an animal can perceive, it has some perspective on its objectives" (2009, p. 267).

One might think that this is the level at which agency truly emerges. This is what Sean Thomas Foran (1997) argues. According to Foran, an animal moves itself, as opposed to being passively moved as a rock is, when the animal's movements are shaped with respect to objects of that

animal's perception. Foran's notion of movement being shaped seems similar to the notion I offered just above, of a system's behavior being coherently produced.

> "Movement shaped with respect to an object of perception" does not simply mean "movement caused by perception." Movement can be caused, in some quite general sense of "caused," by perception without being shaped with respect to the object of that perception. Consider this example. Suppose that when a certain kind of quadruped animal sees one of its natural predators, it immediately lowers itself to the ground and remains still. Perceiving the predator causes the animal to lower itself, but the movement that is caused is not shaped with respect to the predator. The movement is still shaped with respect to something the animal perceives, the ground, but its perception of the ground is not what led it to lower itself: this episode of movement was caused by perceiving the predator.
>
> (1997, pp. 41–42)

At this level, perhaps, it becomes appropriate to think of coherent production of behavior in terms of practical rationality. When a system can represent behavioral targets, and implement plans for behavior that approximate standards of success regarding these targets, that system's behavior might well be considered practically rational. And some of that system's behavior might be considered intentional action.

We are still, however, at a level of relative simplicity. At this level, it is important that the system be embedded in circumstances in the right ways. For, while the system may be able to represent targets for behavior and deploy plans to hit these targets, the behavioral profiles deployed in following the plan may be *inflexible*. And inflexible behavioral profiles contain a flaw regarding the meeting of certain behavioral standards.

Distinguish between success according to the standard a system's particular goal or plan sets, and success according to the standards that apply to that system more broadly. If the system is at all complex, then the standards that apply to it will be broader than the standards a particular goal or plan sets. It will have a range of needs, or perform a range of functions. It may even have a range of intentions, which need to be delicately executed in order not to fail with respect to some of them. Inflexible behavioral routines lock the system into one way of behaving, making it difficult for the system to change tack, or to adjust even slightly. As a result, any infelicitous circumstances, or any kinks in the plan, may throw the system off course.

Consider the digger wasp, *Sphex ichneumoneus*. In preparing to lay her eggs, the Sphex displays some extraordinarily intelligent-seeming behavior. It catches and drags a cricket into its burrow, lays eggs in the burrow, closes the burrow, and leaves.

Regarding this behavior, Woolridge (quoted in Dennett, 1984, p. 11) comments (though the actual details regarding Sphex behavior may be more complicated—see Keijzer, 2013):

> To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness – until more details are examined. For example, the Wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is alright.
>
> (1963, p. 82)

Apparently, the Sphex will do this repeatedly, no matter how many times one tampers with the cricket. Commenting on the Sphex's strange behavior, Dennett writes: "Lower animals, such as Sphex, are constitutionally oblivious to many of the reasons that concern them" (1984, p. 24). By reasons, Dennett is referring to certain courses of action rationalized by the animal's own background needs, drives, and (if such states can be legitimately attributed to the animal) beliefs and desires. One problem with the Sphex's behavior is it appears blind to a wide range of pressing practical reasons, in the sense that the animal can be placed in circumstances that render it systematically poor at achieving its own basic goals.

Now, the range of circumstances in which a system can follow or approximate various behavioral standards will probably vary by degree. In biological creatures, increasingly sophisticated psychological structures correlate with a wider range of behavioral success. And simpler structures correlate with gaps in rational behavior (cf. Hurley, 2003).

For example, the honeybee has evolved a richly combinatorial communicative system—the waggle dance—and a good navigational system. The properties of one honeybee's waggle dance will tell other honeybees where to go to find nectar. But consider a series of experiments in which Gould and Gould (1988) (and Tautz et al. (2004)) had honeybees discover nectar in the middle of the lake, which they then reported to their colleagues. Almost as if they didn't believe what they were seeing, the honeybees ignored the waggle dance. One interpretation of this, as Camp (2009) notes, is that the bees put the states <nectar is there> and <there is lake> together into the state <nectar in lake>, which they subsequently rejected. But an alternative interpretation is that the bees failed to make sense of what they saw because of a limit in their representational system. As Camp puts it, "Perhaps their representation nectar there is blocked from interacting with their cognitive map, because the

region on the map marked 'lake' can't receive any other markers" (2009, p. 299). If that is right, then the bees have a representational limit that renders them unable to accord with the relevant norm in one circumstance, even though their representational system is overall well-tuned to deliver success.

Like the rest of the animal kingdom, human beings have representational and psychological limitations. But unlike most other animals, human beings have capacities to work with their psychological states and representations in various ways. Penn, Holyoak, and Povinelli have argued that a key feature of the human mind is the ability to reinterpret various kinds of available representations "in terms of higher-order, role-governed, inferentially systematic, explicitly structural relations" (2008, p. 127). Tyler Burge is after something similar when he distinguishes between reasoning and critical reasoning.

> A non-critical reasoner reasons blind, without appreciating reasons as reasons. Animals and small children reason in this way … Not all reasoning by critical reasoners is critical. Much of our reasoning is blind, poorly accessible, and unaware. We change attitudes in rational ways without having much sense of what we are doing. Often, we are poor at saying what our reasoning is. Still, the ability to take rational control of one's reasoning is crucial in many enterprises – in giving a proof, in thinking through a plan, in constructing a theory, in engaging in debate. For reasoning to be critical, it must sometimes involve actual awareness and review of reasons …
>
> (2013, p. 74)

At this stage—a stage of adult human sophistication that can involve reflection on our reasons as reasons, and that can involve considerations of relations of reason between our various psychological states, we find a level of agency that Michael Bratman has developed in much detail. We find *planning agents*:

> In support of both the cross-temporal and the social organization of our agency, and in ways that are compatible with our cognitive and epistemic limits, we settle on partial and largely future-directed plans. These plans pose problems of means and preliminary steps, filter solutions to those problems, and guide action. As we might say, we are almost always already involved in temporally extended planning agency in which our practical thinking is framed by a background of somewhat settled prior plans.
>
> (Bratman, 2018, p. 202; see also, Bratman, 1999, 2007)

As Bratman's work makes clear, planning agents are veritably bathed in applicable practical norms. We spend much of our time working through

implications of our commitments, testing them against other possible commitments, wondering whether some other course of action might be better in some way, wondering how the plan will impact others, or whether refinements to the plan might make profit along some unforeseen dimension.

There is an interesting series of correlations. As a system capable of behavior (and action) increases in complexity, so do the practical and rational norms that apply to it. So, then, do the chances of internal practical conflict. So, then, does the value to that system of ways of working through, seeking to avoid, seeking to find resolutions to, existing and potential conflict.

It remains to apply this picture of agency to the notion of a disappearing agent and to draw lessons.

## 1.3 The agent appears

Although agentive complexity, and arguably internal rational conflict, are present before the agentive sophistication of adult humans, this is the level that motivates Velleman's and Wallace's articulation of the disappearing agent problem. We are now in a position to see how this problem afflicts agents. Let us momentarily step back and summarize the discussion.

According to the picture under development, the agent is essentially an integrated system of internal activity and behavioral control that warrants the application of behavioral standards, and contains the capacity or capacities to coherently meet or approximate some sufficient set of these standards. In psychological agents, the standards come to be characterized as rational standards, and the activity that leads to coherent behavioral control comes to be characterized as practically rational activity—often, practical reasoning—and action.

At the same time, in psychological agents like humans, the very complexity of the system that is the agent will often lead to internal rational conflict. This is because of the multiple behavioral and rational standards that will apply to such a system in many circumstances. In cases of internal rational conflict, the agent may seem to disappear. This is because in such cases some important features of the agent are at cross-purposes, and if we wish to answer how the agent is implicated in the action, we will struggle to find a good answer.

Now that we have a better handle on what agents are, however, we can see how the problem of the disappearing agent is no real problem, even though it arises from a truth about the nature of some sophisticated agents, like human beings. The problem is no real problem in the sense that it neither challenges any particular causal theory of action, nor does it motivate the posit of esoteric states or causal capacities. The problem is just a function of the ways human behavior and action is

produced—namely, by an imperfectly organized set of mechanisms and capacities that sometimes end up at cross-purposes, undermining agential unity or agential rationality. Agents often produce behavior and action that is sub-optimal in one or many respects. This does not make the agent disappear. To think otherwise is to give the agent, qua agent, far too much credit.

## 1.4 Mental action as rational glue

I have said that as a system capable of behavior (and action) increases in complexity, so do the practical and rational norms that apply to it, as well as the chances of internal practical conflict, as well as the value to that system of ways of working through, seeking to avoid, seeking to find resolutions to, existing and potential conflict.

In human agents, mental actions—imagination, the direction of attention, counterfactual rumination, attempts to remember, etc.—are one of the main ways we have of navigating internal practical conflict. Other philosophers have noted a close connection between mental action and an agent's capacity to satisfy or otherwise display sensitivity to various applicable norms (Metzinger, 2017; Proust, 2013). The picture in play is one on which many—even if not all – mental actions concern inwardly directed activities aimed at rationality-relevant states, attitudes, and conditions. So, Thomas Metzinger claims that "Mental action is a specific form of flexible, adaptive task control with proximate goals of an epistemic kind: in consciously drawing conclusions or in guiding attention there is always something the system wants to know, for example, the possibility of a consistent propositional representation of some fact, or the optimal level of perceptual precision" (2017, p. 3). When engaging in mental action, an agent is sometimes searching for information, sometimes assessing sets of attitudes for coherence or consistency, and sometimes exploring potential consequences of behavior in light of existing beliefs and desires.

So, the picture of agency developed here makes sense of the compelling thought that mental action is tied to human agency (and to agents with similar psychological structure) in an intimate way. I think the pervasiveness of mental action in our mental lives is a product of our particular computational, informational, and cognitive architectural limitations, as well as the solutions evolution seems to have bequeathed to us. I cannot argue the point in full here, but it seems to me that much of our mental action—and especially the actions that contribute to processes of practical deliberation – is driven by uncertainty and conflict (Shepherd, 2015). This uncertainty and conflict are related to our sense of the norms of practical rationality. Often, in deliberation, we are engaged in a search to uncover what it is best to do, or how best to execute a pre-existing intention, or how best to navigate a conflict between

various desires, or obligations, or commitments, or whatever. We deliberate because we are informationally fragmented in certain ways (Egan, 2008)—it is a struggle to call to mind the relevant items and to put them together in rationally satisfying ways.

Velleman and Wallace were heading in this direction, emphasizing various mental actions by which agents more closely approximate the standards of reason. The error is in suggesting that these mental actions should be embedded into the essence of an agent.[6] That suggestion, I have already said, gives human agents too much credit. For a nearly perfect agent may have little need of the mental actions via which humans rationally glue together their many plans and preferences and aspects of identity.

To see what I mean by this, consider a being constitutively incapable of uncertainty or conflict: an omniscient, omnipotent, and fully practically rational being. Call it Al. It is certainly conceivable that Al, in virtue of its supreme knowledge, never faces uncertainty. And Al, in virtue of its full practical rationality, never faces conflict (unless it be a conflict in the very norms of practical rationality). In whatever the situation, no matter how complex, Al need not deliberate to discern the best course of action. We might say that no matter the situation, no matter how complicated or fraught with moral gravity, Al simply sees—takes in at a glance—the thing to do. Al always acquires intentions reflective of Al's omniscience and full practical rationality. (In order to take in all the information required to always discern the thing to do 'at a glance,' Al will need some pretty amazing perceptual sensitivity and some pretty amazing cognitive sophistication. We can assume this is covered by Al's omnipotence.)

It seems to follow that neither the kind of uncertainty and conflict that is our normal situation, nor the actional processes of deliberation and decision via which we attempt to reduce uncertainty and accord with norms of practical rationality, are essential for agency (Arpaly and Schroeder (2012) and Morton (2017) make the same point). Further, it seems to follow that uncertainty, deliberation, and decision are important features of our—that is, human—agency precisely because human agency is far from perfect. We have perceptual, cognitive, and volitional limitations, and it is because of this that uncertainty, deliberation, and decision play such a large role in our lives.

Even if these kinds of mental actions are inessential to the nature of agency, for agents like humans, the activity of practical reasoning that is essential to our agency is often conducted via mental actions—intentional mental activities like shifts of attention, inhibition of urges, imagination of possibilities for action or consequences of courses of behavior, comparison of action options, weighing of reasons, and so on. I want to suggest that for human agency, mental action is a rational response to the computational, informational, and architectural limitations we

face. Mental action is a kind of rational glue—it is one key way that we attempt to discover the norms of practical rationality, and to enforce rational coherence across the large but disjointed set of goals, preferences, and abilities that we tend to possess.

## Notes

1. Research for this chapter was funded by the European Research Council's Horizon 2020 program, Starting Grant ReConAg 757698.
2. This chapter was written over a period of time during which I was also writing a book, *The Shape of Agency* (Shepherd, 2021). Chapter 5 of that book is devoted to an exploration of the nature of agency. My end-game there is different than here, but both that chapter and this one required development of my thinking regarding the nature of agency. So, the ideas in that book and this chapter bear relations of mutual influence to one another. As a result, what I say in §2 of this chapter extracts, in some cases repeats, and in other cases reformulates, parts of that chapter.
3. The problem of the disappearing agent that I discuss here is thus distinct from Derk Pereboom's (2014) presentation of a problem for event-causal libertarian views of free will. I mention this because Pereboom calls the problem he discusses the problem of the disappearing agent. See Randolph Clarke (2019) for a penetrating discussion of Pereboom's argument, and of disappearing agent considerations more generally.
4. The dialectic here is a little dirty, since Velleman is not arguing that disappearing agent cases are not cases of action. He is interested, instead, in a notion he calls action par excellence—the exemplification of full-blooded agency. Event-causalist views are not charitably taken as attempting to capture this notion (see Mele, 2003). So, it seems Velleman is better read here as raising a deeper question about the nature of agency, as opposed to a specific problem for event-causal views of intentional action.
5. The same need can be seen as motivating earlier proposals—i.e., Frankfurt's (1988) involving 'identification' and Watson's (1975) involving the agent's system of values.
6. Again, to be fair to Velleman, he embeds this into the essence of agency par excellence (see also Footnote 4). But it is reasonable to be uncertain about the fruitfulness of this category (see Mele, 2003).

## References

Arpaly, N., & Schroeder, T. (2012). Deliberation and acting for reasons. *Philosophical Review*, *121*(2), 209–239.

Brand, M. (1984). *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.

Bratman, M. (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge, MA: Cambridge University Press.

Bratman, M. E. (2007). *Structures of agency: Essays*. Oxford: Oxford University Press.

Bratman, M. E. (2018). *Planning, time, and self-governance: Essays in practical rationality*. Oxford: Oxford University Press.

Brent, M. (2017). Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy*, *47*(5), 656–673.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, *79*(2), 251–278.
Burge, T. (2013). *Cognition through understanding: Self-knowledge, interlocution, reasoning, reflection: Philosophical essays*, *volume 3*. Oxford: Oxford University Press.
Camp, E. (2009). Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research*, *78*(2), 275–311.
Clarke, R. (2019). Free will, agent causation, and "disappearing agents." *Noûs*, *53*(1), 76–96.
Davidson, D. (2004). *Problems of rationality*. Oxford: Clarendon, Oxford University Press.
Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.
Egan, A. (2008). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47–63.
Foran, S. T. (1997). *Animal movement*. University of California, Los Angeles. ProQuest Dissertation Publishing.
Frankfurt, H. G. (1988). *The importance of what we care about: Philosophical essays*. Cambridge: Cambridge University Press.
Gould, J. L., & Gould, C. G. (1988). *The honey bee*. New York, NY: W. H. Freeman.
Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–257.
Keijzer, F. (2013). The Sphex story: How the cognitive sciences kept repeating an old and questionable anecdote. *Philosophical Psychology*, *26*(4), 502–519.
Kung, C., & Saimi, Y. (1982). The physiological basis of taxes in Paramecium. *Annual Review of Physiology*, *44*(1), 519–534.
List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.
Mele, A. R. (1992). *Springs of action: Understanding intentional behavior*. Oxford: Oxford University Press.
Mele, A. R. (2003). *Motivation and agency*. Oxford: Oxford University Press.
Metzinger, T. (2017). The problem of mental action: Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing: 19*. Frankfurt am Main: MIND Group. DOI: 10.15502/9783958573208.
Morton, J. M. (2017). Reasoning under scarcity. *Australasian Journal of Philosophy*, *95*(3), 543–559.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109–130.
Pereboom, D. (2014). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies*, *169*(1), 59–69.
Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford: Oxford University Press.
Shepherd, J. (2015). Deciding as intentional action: Control over decisions. *Australasian Journal of Philosophy*, *93*(2), 335–351.
Shepherd, J. (2021). *The shape of agency: Control, action, skill, knowledge*. Oxford: Oxford University Press.

Tautz, J., Zhang, S., Spaethe, J., Brockmann, A., Si, A., & Srinivasan, M. (2004). Honeybee odometry: Performance in varying natural terrain. *PLoS Biology*, *2*(7), e211.

Velleman, J. D. (1992). What happens when someone acts. *Mind*, *101*(403), 461–481.

Wallace, R. J. (1999). Three conceptions of rational agency. *Ethical Theory and Moral Practice*, *2*(3), 217–242.

Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*(8), 205–220.

Wooldridge, D. E. (1963). *The machinery of the brain*. New York: McGraw-Hill.

# 2 How to Think Several Thoughts at Once

## Content Plurality in Mental Action[1]

*Antonia Peacocke*

### 2.1 Introduction

Some mental actions are intentional. Some of these are performed by way of performing another kind of intentional action (i.e. as a **constitutive means**). Just some of these have a striking and philosophically important feature: they have their own contents, but they are performed by way of mental actions with distinct contents. These mental actions have **content plurality**.

Here are a few examples. You can judge that *this laundry detergent is the cheapest* as a constitutive means to deciding *I'll buy this laundry detergent*. Similarly, the mental action of judging *25 + 34 = 59* can be your means of judging *my meal costs $59 in total*. Your calling the word *"phlogiston"* to mind can also constitute a decision that *"phlogiston" will be the treehouse password*. The latter action in each of these descriptions is a **complex mental action**: an intentional mental action executed by doing another kind of thing intentionally. But each such mental action also has a further important feature. Each is a mental action with some specific content performed by way of a mental action with a distinct content. In other words, each such complex mental action has **content plurality**.

The fact that mental actions can have content plurality has not previously been appreciated in philosophy. If it had been, several philosophical debates of the last few decades would have run considerably differently. The content plurality of mental actions transforms several debates: one about the epistemology of transparent self-knowledge; one about the relationship between doxastic judgments and practical decisions; and one about the nature of inference. In each such debate, the content plurality of mental actions opens up a new and attractive solution to a philosophical puzzle that has not been considered before.

Here is a brief outline of this chapter. In Section 2.2, I identify the kind of mental actions that have content plurality, and give some examples. In Section 2.3, I give five jointly sufficient conditions on a mental action's

DOI: 10.4324/9780429022579-3

having content plurality, and I explain how it is possible to meet these conditions together. In Section 2.4, I use this model of content plurality in mental action to advance the three distinct philosophical debates I just mentioned. In Section 2.5, I respond to two objections to the view developed here.

## 2.2 Mental actions with content plurality

I'll start by identifying the kind of mental actions that have content plurality. I'll do that in a series of steps that narrows down mental actions into more and more restrictive categories.

Here are the steps in outline form:

1 There are mental actions.
2 Some of those are intentional mental actions.
3 Some of those are intentional mental actions that have contents.
4 Some mental actions are performed to satisfy more than one intention at once.
5 Some of those are mental actions that **execute** several intentions at once.[2]
6 Some of those mental actions execute several intentions with **content conditions**.
7 Some of those execute one such intention by having one content, and execute another such intention in constituting another kind of action with a distinct content. Any such further action has **content plurality**.

Let's take each of these narrowing steps individually.

1 There are mental actions.

A mental action is something you do in thought. You can do all sorts of things in thought. You can add 25 and 34. You can recall a recent party. You can rehearse a lecture you will give. You can imagine a firework show. You can do any of these things without putting a pen to paper, speaking, gesticulating, or even moving your body in any way. It's not necessary for you to sit completely still to do something mentally, but nor is it necessary to move your body at all.

2 Some of those (mental actions) are intentional mental actions.

Some mental actions are intentional. You can do any of those I just mentioned intentionally. You can intentionally add 25 and 34. You can recall a recent party intentionally. You can intentionally rehearse a lecture, or intentionally imagine a firework show.

That does not mean that any kind of mental action you can perform is a kind of mental action you can perform intentionally. There are two relevant constraints. First, some descriptions of mental actions could not figure into an intention, so you cannot do something intentionally in thought under any such description. Second, there are mental actions you could try to do intentionally, but you will necessarily fail.

To illustrate this first constraint, here is an example of an intention you can't have. You cannot take some proposition you overtly take to be false – say, that 60 divided by 3 is 10 – and intend to make a judgment with its content. You cannot have that intention, because it would require using the concept JUDGMENT, and to have that concept is also to understand that judgment involves commitment to the truth of a proposition.[3] Not only can you not judge something you overtly take to be false; you also know you cannot do that. And knowing you cannot do something is incompatible with intending to do it. The point is not just that you will not succeed if you try to do it; it's that you cannot have the intention to do this at all.[4]

To illustrate the second constraint, here is an example of an intention you cannot successfully execute. You might act on an intention not to think about a polar bear, thereby trying not to think of a polar bear. But insofar as you are directly acting on that intention, you are thinking of what you are doing under the same description under which it is intentional. That is, you're thinking of what you're doing as *not thinking about a polar bear.*[5] But to do that is to think about a polar bear.[6]

Due to these two constraints, you cannot perform just any kind of mental action intentionally. But there are many kinds you can and do perform intentionally.

There are difficult unsolved puzzles about what it is to act intentionally. I will not attempt to solve them here. The main claim of this chapter – that mental actions can have content plurality – does not hang on any one particular solution to these difficult puzzles.

What will be important is the fact that any intentional mental action is intentional *under a description*, and not every true description of a mental action is also a description under which that action is intentional.[7] Thus, in intentionally *adding 25 and 34*, you might in fact be doing precisely what your classmate is doing at the same time. While your *adding 25 and 34* is intentional under that very description, it is not also intentional under the description *doing what your classmate is now doing* – even though it is indeed an instance of doing just that.

3 Some of those are intentional mental actions that have contents.

Some intentional mental actions have contents under the descriptions that characterize them as intentional. When you imagine a firework show intentionally, your imagining has firework show content – perhaps

an imagistic content. When you intentionally recall a party, your recollection is of that party. When you intentionally rehearse a lecture you will give, you think of various sentences you will speak aloud. When you (correctly) add 25 and 34, you make a judgment with the propositional content *25 + 34 = 59*. Intentional mental actions can have all sorts of different kinds of content – at least imagistic, propositional, and linguistic contents.

It may sound odd to say that actions can have contents. And it is true that not all actions have contents. But you might think, additionally, that an action isn't the right kind of thing to have a content, and that there is a category error in statement (3) above.

To dispel this confusion, it is important to see that the category of action is a determinable category, of which there are many determinates. An action can be a kick, a heist, an election, or something else. A mental action can be a judgment, a decision, a recollection, and imagining, or something else. These more determinate categories of actions can clearly have contents. Take the category of judgment: a judgment must have a content to be a judgment at all. Similarly, each decision has a content, since each decision is a decision to do something.

To be most precise, you might insist that intentional mental actions have contents only under certain descriptions. But since these are descriptions that really do apply to these actions, it seems fair to say (as I will) that intentional mental actions themselves have contents.

I will also use another piece of terminology that will help us frame the issues of this chapter: an intentional mental action has content under one of its intentional descriptions just when the corresponding intention specifies a **content condition** to be met. For example, your intentional *imagining of a firework show* is a case in which you act on an intention that demands an episode of imagining with some firework show content. If you succeed at imagining one, your action meets the content condition specified by the intention on which you act.

There is an important caveat here. Even those intentions with content conditions do not always fully specify the most determinate content of the mental action to be performed. The opposite is usually the case: the most determinate content of your action is more determinate than the content specified by the intention on which you act in performing that action.

For example, when you intentionally add 25 and 34, your intention is *to add 25 and 34*. In this case, you intend to judge, of some particular number (de dicto), that it is the sum of 25 and 34. What you do in order to execute this intention is to figure out which number that is. The ultimate content of the action that executes the intention with which you began is the proposition *25 + 34 = 59*. This content meets the content condition with which you began in part because it is more determinate than the specification built into the intention.

Take another example, this time from the practical domain: an example of intentionally deciding where to go for lunch. When you do this, you do not start with a specific restaurant in mind (de re); but the successful execution of this intention demands a specific restaurant (de re). This is another case in which the determinate content of your successful mental action is more determinate than was specified by the content condition built into your intention.

4 Some mental actions are performed to satisfy more than one intention at once.

An action can be performed to satisfy more than one intention at once.[8] This is the case when you intentionally *Φ in order to Ψ*, for two action types Φ and Ψ.[9] To do this is (inter alia) to take Φ-ing as at least a partial means to Ψ-ing. For instance, when you intentionally *turn on the oven in order to make dinner*, you take *turning on the oven* to be part of *making dinner*. Your turning on the oven is, in this case, intentional under the description *turning on the oven*; it is done intentionally under that description in order to *make dinner* intentionally under that further description; and if all goes well, your intentional action of *turning on the oven* will partially constitute your intentionally *making dinner*. Here, your intentionally *turning on the oven* is a **partial means** to *making dinner*. Your intentionally *making dinner* here is a **complex action**, one performed by means of doing other kinds of things intentionally (including, but not limited to, your turning on the oven).

There are also cases like these involving mental action instead of bodily action.

When you intentionally *Φ in order to Ψ*, Φ-ing is sometimes only a partial means to Ψ-ing. Here is an example of a mental action like that. You might *imagine the taste of the fettucine dish in order to choose an entrée*. In this case, your imagining of the taste of the fettucine dish is intentional under that very description. It is intentional under the description *imagining the taste of the fettucine dish*; you do that intentionally in order to *choose an entrée* intentionally under that further description; and if all goes well, your intentional action of *imagining the taste of the fettucine dish* will partially constitute your intentionally *choosing an entrée*. It only partially constitutes that further intentional action because just imagining one dish is not clearly sufficient for fully choosing an entrée. It might be only one step along the way to your choice. Here, as before, your intentionally *imagining the taste of the fettucine dish* is a partial means to *choosing an entrée*. Your intentionally *choosing an entrée* is a complex action, one performed by means of doing other kinds of things intentionally (including, but not limited to, your *imagining the taste of the fettucine dish*).

Here is another example, this time with a mental action with propositional content. You can intentionally *suppose that p* [for some determinate *p*] *in order to figure out whether p only if q* [for some determinate *q*]. Your action is intentional under the description *supposing that p*; you do that intentionally in order to *figure out whether p only if q* intentionally under that further description; and if all goes well, your intentional action of *supposing that p* will partially constitute your intentionally *figuring out whether p only if q*. It only partially constitutes that further action because just supposing that *p* is not, alone, a way of *figuring out whether p only if q*. After supposing that *p*, you have more left to do in order to figure out whether *p only if q*.

5 Some of those are mental actions that **execute** several intentions at once.

Sometimes, you execute two or more intentions in just one token mental action. In some cases when you intentionally *Φ in order to Ψ*, your Φ-ing is not just one of many steps on the way to Ψ-ing. Instead, your Φ-ing can also fully constitute your Ψ-ing. In such a case, your Φ-ing intentionally is a **constitutive means** to your Ψ-ing intentionally. Your Ψ-ing is a complex action performed just by Φ-ing intentionally.

In such cases, you execute your intention to *Φ* and your intention to *Ψ* all at once. You **execute** an intention to *Φ* just when you Φ intentionally in acting on that very intention. We can also say that an action is the **execution** of an intention just when that action is itself a Φ-ing performed intentionally in acting on that intention.

Let's look at some examples to make all of this more concrete.

First, for comparison, here is a bodily action that executes more than one intention at once. I can intentionally *knit a scarf in order to make a gift for my brother.* If I succeed in knitting the scarf intentionally (in acting on this very intention), I also thereby succeed in making a gift for my brother. Here, I intentionally *knit a scarf*, under that description; I do that intentionally in order to *make a gift for my brother*; and that successful intentional *knitting a scarf* also fully constitutes an intentional *making of a gift for my brother.* This time, the intentional action of *knitting a scarf* executes the intention to *knit a scarf* as well as the intention to *make a gift for my brother.* I can execute all these intentions at once because there is no more I need to do to make a gift for my brother other than *knit the scarf* in these circumstances. (I am here assuming that gift wrap and a card are supererogatory.)

That was an example in the domain of bodily action. More importantly for our purposes, there are also mental actions that execute more than one intention at once. Here is an example. If the teacher of a meditation class asks you to attend to the sound in the room, you might form an intention to *attend to the sound in order to do what the meditation*

*teacher asked*. If you then go on to attend to the sound, that intentional mental action can execute your several intentions at once: your intention to *attend to the sound* as well as your intention to *do what the meditation teacher asked*. Attending to the sound can execute all that at once because there is nothing more you need to do – over and above attending to the sound – in order to do what your meditation teacher asked in these circumstances.

6 Some of those mental actions execute several intentions with **content conditions**.

We can now combine point (3) above – the point that there are intentional mental actions that have contents – with the point that a mental action can execute several intentions at once.

I'll say that an intention has a **content condition** just if any action that executes it must have a certain kind of content.

We have already considered an example of a mental action intentional under one description that also partially constitutes an intentional action of another kind. In this example, the relevant intentions involved content conditions. This was the example of *supposing that p in order to figure out whether p only if q*. But this is not a case in which you execute, all at once, several intentions that have content conditions. That is because supposing that *p* is not, by itself, a fully constitutive means of figuring out whether *p only if q*.

Now consider a case in which one action executes several intentions with content conditions – indeed with distinct content conditions. If you want to meet your friend Ben for dinner out, you can intentionally *choose a restaurant in order to choose a place to meet Ben*. If you successfully *choose a restaurant* while acting on this intention, there is nothing left to do to choose a place to meet Ben. Your intentional action of *choosing a restaurant* also fully constitutes your intentional action of *choosing a place to meet Ben*. (How this can be so is a question left for the next section of this chapter.)

This is a case in which one mental action executes several intentions at once, each of which builds in a content condition. In other words, it is a case of a complex contentful mental action performed via some constitutive means which is itself a contentful mental action. However, this complex action is not yet a case of a mental action with content plurality. It does not have content plurality because the action that is the constitutive means has the same content as the further complex action. Your intentional *choosing of a restaurant* has a specific content – let's say, the local restaurant Taïm. Your intentional *choosing of a place to meet Ben* also has the content Taïm. The restaurant you choose is Taïm, and the place you choose to meet Ben is Taïm. This is a case in which two intentions with distinct content conditions – one demanding a restaurant as

such, one demanding a place to meet Ben as such – can both be executed at once because the content of the action that is your constitutive means is one and the same as the content of your further action.

7 Some of those execute one such intention by having one content, and execute another such intention in constituting another kind of action with a distinct content. Any such further action has **content plurality**.

In some cases, you execute several intentions with distinct content conditions at once, by using distinct contents to meet those conditions. You can do that when you intentionally perform one kind of contentful mental action as a constitutive means to another, in such a way that the means action has a content distinct from that further complex action it constitutes. Such complex actions just are mental actions with **content plurality**.

Here is an example. Say you have just enjoyed a meal consisting of a $25 appetizer and a $34 entrée. In this context, you can intentionally *add 25 and 34 in order to determine the total cost of the meal*. When you successfully *add 25 and 34*, you make a judgment with the content *25 + 34 = 59*. When you successfully *determine the total cost of the meal*, you make a judgment with the content *the total cost of the meal is $59*. You can execute both intentions at once. Your intentional action of *adding 25 and 34* has the content *25 + 34 = 59*. In context, used as a constitutive means to *determining the total cost of the meal*, that intentional action also constitutes an intentional action of that further type. That is, it also constitutes a mental action of *determining that the total cost of the meal is $59*. That complex mental action has a different total content than that of the constitutive means; it has the content *the total cost of the meal is $59*. Since it was performed by way of performing another kind of intentional mental action with a distinct content, that action of *determining the total cost of the meal* has content plurality.

To give a preview of the next section: the reason that you can execute both of these intentions at once is that you already understood, in entering into this mental activity, that the total cost of the meal in dollars would be the sum of the cost of the appetizer and the cost of the entrée, i.e., the sum of 25 and 34. Although the content conditions built into the intentions on which you are acting are themselves distinct – one intention demands just a mathematical judgment, and another intention demands a judgment about your meal – your background understanding relates them in a way that makes them executable at once.

Here is another example. Say you and your life partner build a treehouse in the backyard, and you need to set a password for entry to the treehouse. One way to do this is just to *call a word to mind in order to decide on a treehouse password*. This is another case in which

you can execute both your intentions at once, although each intention sets a distinct demand on the content of the mental action that would execute it. You can execute both intentions at once because when you intentionally *call a word to mind*, there is no more you need to do to *decide on a treehouse password*. One mental action can both execute your intention to *call any word to mind* by having a linguistic content, e.g., "phlogiston" (the word), and also execute your intention to *decide on a treehouse password* by having the content *"phlogiston" will be the treehouse password*. These are contents of distinct kinds. One is linguistic ("phlogiston") and one is propositional (*"phlogiston" will be the treehouse password*). Your password decision here is performed by way of your performing another kind of mental action with a distinct content—the kind of simple linguistic content that couldn't even in principle be the content of a decision. Your password decision thus has content plurality.

Let's take stock. So far I have simply identified mental actions with content plurality among the more general class of mental action. I have provided a few different examples.

However, I haven't explained how such mental actions are possible. That is the next task.

## 2.3 How content plurality is possible

What does it take for a mental action to have content plurality? In this section, I'll state jointly sufficient conditions for that, and then explain how it is possible to meet these conditions.[10]

### *2.3.1 Jointly sufficient conditions*

Here are jointly sufficient conditions on performing a mental action with content plurality.

For two distinct types of contentful mental action Φ and Ψ, if

i you think that Φ-ing is a way to Ψ in your circumstances, because Φ-ing bears a certain relation *r* to Ψ-ing in your circumstances;
ii you act on an intention to *Φ in order to Ψ*, led by this conception of Φ-ing;
iii all it takes to Ψ in your circumstances is to think of a token Φ-ing of yours as bearing that same relation *r* to Ψ-ing, and
iv you execute both intentions (to *Φ* and to *Ψ*) just by Φ-ing intentionally in such a way that
v the content of your Φ-ing is qualitatively distinct from the content of your Ψ-ing,

then your Ψ-ing has content plurality.

For an example of a mental action that meets all these conditions, return to the restaurant math case. Let Φ be *adding 25 and 34* and Ψ be *determining the total cost of the meal*. In this case,

i you think of *adding 25 and 34* as a way to *determine the total cost of the meal* in your circumstances, because the result of adding 25 and 34 is the total cost of the meal;
ii you act on an intention to *add 25 and 34 in order to determine the total cost of the meal*, led by the aforementioned conception of those actions;
iii all it takes to *determine the total cost of the meal* in your circumstances is to think, of the result of *adding 25 and 34*, that it is the total cost of the meal (as indeed it is); and
iv you execute both intentions (to *add 25 and 34*, and to *determine the total cost of the meal*) just by adding 25 and 34 intentionally in such a way that
v the content of your *adding 25 and 34*—that is, the propositional content *25 + 34 = 59*—is qualitatively distinct from the content of your *determining the total cost of the meal*, which is the propositional content *the total cost of the meal is $59*.

Here, your intentional action of determining the cost of the meal has content plurality.

### *2.3.2 An explanation of content plurality*

To accept (i) – (v) as jointly sufficient conditions on a mental action's having content plurality, however, is not yet to accept that there really are mental actions with content plurality. To accept that, we need to see how it is possible to meet conditions (i) – (v) together.

Here's how I'll go about explaining that. I'll work backwards from the end. First I'll establish that the conclusion—that the complex mental action Ψ-ing—follows from (i) – (v). I'll do that by assuming (i) – (v) and establishing the conclusion.

In the following step, I'll assume (i) – (iv), and show that (v) is possible given these. If so, this makes content plurality possible as well. Then I'll assume just (i) – (iii) and show it's possible to meet the other conditions; then do the same starting just with (i) and (ii), and then just (i). In the end, our sole remaining task will be to show that meeting (i) is possible. Altogether, establishing these points will establish that it is possible to meet (i) – (v) together.

First, let's establish that if (i) – (v) hold for two action types Φ and Ψ, then Ψ-ing has content plurality. Note that your Φ-ing executes both relevant intentions (to *Φ* and to *Ψ*) at once, as given by (iv). That means that your Φ-ing also constitutes a Ψ-ing. Since (v) tells us that

your Φ-ing has a qualitatively distinct content from your Ψ-ing, your Ψ-ing is a mental action with content plurality: the constitutive means by which it is performed has a content qualitatively distinct from its own content. This establishes that your Ψ-ing's having content plurality follows from conditions (i) – (v), since it is assumed that there is nothing special about *A*.

Now let's see how it is possible to meet all these jointly sufficient conditions together.

Given that conditions (i) – (iv) can be met, can (v) be met as well? We have already seen that one action of *adding of 25 and 34* can also constitute an action of *determining the total cost of the meal*. But one of those is just a judgment about the sum of 25 and 34, and the other is a judgment about the total cost of the meal. As such, then, they must have distinct contents. If they can be executed all at once—as condition (iv) stipulates—then that action must have content plurality. Other examples from the first part of this chapter give ample illustration of the ways in which condition (v) can be met given that (i) – (iv) are met.

Given that conditions (i) – (iii) can be met, can (iv) be met as well? This is the trickiest point to see. Explaining this point will take some time, and use several principles from the philosophy of action more generally. I will take the explanation step by step.

Condition (iv) stipulates that you execute several distinct intentions – to *Φ*, to *Ψ*, and to *Φ in order to Ψ* – all at once, in one token mental action. How can this be?

Consider conditions (i) and (ii) first. Condition (ii) stipulates that you act on a pair of intentions that relates Φ-ing and Ψ-ing as means to end. Here, you are trying to *Φ in order to Ψ*. Sometimes when you act on a complex intention like this, you see Φ-ing as only a partial means to Ψ-ing. But (i) gives us more than that. It stipulates that you see Φ-ing as a way to Ψ in your circumstances. By that, I meant that you consider Φ-ing to be a sufficient means to Ψ-ing in your circumstances. In acting on this intention, you think that in Φ-ing, you will also Ψ. You think this because you think Φ-ing bears some particular (de re) relation *r* to Ψ-ing in your circumstances.

Now we need some further principles about action more broadly. First, acting on an intention to *Φ* necessarily involves having *Φ-ing* as such in mind. The way this point is traditionally put is like this: in acting on an intention to *Φ*, you have in mind what you're doing under a description – the same one under which your acting is intentional.

To say you have this in mind is to say that your conception of what you're doing is present for you in an immediate way. It's not just that you have some standing belief about your current action. What you have in mind in this way has been called "practical knowledge" by Anscombe.[11] I will call it a "practical conception," thereby not yet committing to the truth or warrant of this conception of what you are doing.[12]

Having any such practical conception in mind involves thinking that what you are now doing is a certain sort of activity – namely, *Φ-ing*. This is a form of doxastic commitment.[13] A practical conception present in intentional action involves the same attitudinal aspect as a standing belief or an occurrent judgment. But having a practical conception in mind while you act is neither just to have a belief (because it is by definition immediately present to mind) nor just to make a judgment (because your having this conception in mind extends over time).

In the case of complex actions—cases in which you act on an intention to *Φ in order to Ψ*—the practical conception you have in mind is even richer than that. In such cases, your practical conception involves having in mind: *I am Φ-ing in order to Ψ*. To think this is also to think that Φ-ing is at least a partial means of Ψ-ing, and sometimes—as when condition (i) holds—also to think that Φ-ing is a way, or a sufficient means, of Ψ-ing. (If you didn't think Φ-ing was at least a partial means of Ψ-ing, you wouldn't be *Φ-ing in order to Ψ* at all.)

This understanding of Φ-ing as a way of Ψ-ing doesn't come from nowhere. Whenever you act on a complex intention to *Φ in order to Ψ*, your practical conception also enfolds a more determinate relation *r* between Φ-ing and Ψ-ing—a relation between these two types of action which explains why one is a means to the other, at least in your current circumstances.

For an illustration of this fact, let's return once again to the restaurant math example. As Φ-ing here is *adding 25 and 34*, and Ψ-ing is *determining the total cost of the meal*, then acting on an intention to *Φ in order to Ψ* involves commitment to Φ-ing's being a way of Ψ-ing because the result of adding 25 and 34 is the total cost of the meal in dollars. This is just one example of a more determinate relation that explains your thinking that Φ-ing is a way of Ψ-ing in your circumstances. The particular relation between Φ-ing and Ψ-ing that you have in mind as part of your practical conception—the one which explains why Φ-ing is a way of Ψ-ing in your circumstances—will vary from case to case, of course.

Now we can put these points to work to answer the question at hand. We are asking: given conditions (i) – (iii), how is it possible for (iv) to be met too?

If it is possible to Φ at all—as indeed in some cases it will be—it is possible to execute the intention to Φ in one mental action. Let's assume that you do that.

We have just seen that when you do that you already have in mind, as part of your practical conception of what you are doing, that Φ-ing in your circumstances bears that relation *r* to Ψ-ing, and (because of this) that Φ-ing is a way of Ψ-ing in your circumstances. Your practical conception involves relating a Φ-ing in your circumstances (de dicto) to a Ψ-ing in your circumstances (de dicto). When you actually Φ, this

practical conception is enriched into a conception of your actual token Φ-ing, de re. Thus, in Φ-ing, you think that your Φ-ing (now de re) is related to Ψ-ing in the way built into your practical conception of what you are doing. But that just is to think of a token Φ-ing of yours in precisely the way that makes it into a Ψ-ing as well, as stipulated by condition (iii). Condition (iii) says that all it actually takes to Ψ in your circumstances is to think (de re) of some token Φ-ing you perform as bearing that same relation *r* to Ψ-ing that is built into your practical conception. Thus, your actual token Φ-ing also constitutes a Ψ-ing in these circumstances, by condition (iii). If you have acted intentionally, and non-deviantly, keeping in mind this practical conception all along, then condition (iv) is fulfilled too. Thus, it is possible to meet condition (iv) when conditions (i) – (iii) are met.

To illustrate all this, let's return yet again to your math in the restaurant. In *adding 25 and 34 in order to determine the total cost of your meal*, you have in mind a practical conception of what you're doing under that very description. In having this conception in mind, you think that *adding 25 and 34* is a way of *determining the total cost of your meal* – here, because the result of adding 25 (the cost of your appetizer in dollars) and 34 (the cost of your entrée in dollars) will simply be the total cost in dollars of the meal that you want to figure out.

In summary, you have in mind, as part of your practical conception of what you are doing, that *25 + 34 is the total cost of your meal in dollars*. This simplification looks just like a relation between certain contents, but it's crucial that this relation on the content level is being used by you in thought to perform actions with the relevant contents. You think of the result of your addition to be performed as the total cost of the meal in dollars to be figured out.

Now we can see what happens when you successfully add 25 and 34—which is indeed possible to do—and come to recognize that *25 + 34 = 59*. When you do this, you are already thinking that the sum in this content is *the total cost of the meal in dollars*. This is not a matter of merely entertaining a relationship between the two; you are doxastically committed to this relation. And this is not some standing or unconscious belief you have to call to mind to use. This relationship is present in your practical conception; it makes sense of what you are doing as *adding 25 and 34 in order to figure out the total cost of your meal*. Because your thought that *(25 + 34)* is *the total cost of your meal in dollars* is present to mind in this way—and, as stated in condition (iii), there's nothing more to *figuring out the total cost of the meal* than thinking of this sum in this way—your adding *25 and 34* to get *59* just is also a judgment that *the total cost of your meal is $59*. There is nothing extra you have to do to make that second judgment here.

We have now seen that, given (i) – (iii), it is possible to meet condition (iv) as well. That is not guaranteed; after all, you could simply get

distracted, or otherwise fail to execute any of the intentions on which you are acting. But it is certainly possible, and that is all we need.

Given (i) – (ii), is it possible for (iii) to be met? Here, once again, we can refer back to the examples from Section 2.1. There are certainly mental actions Φ and Ψ such that all it takes to Ψ, in some circumstances, is to Φ while thinking of your token Φ-ing in a particular way. All it takes to *decide on a treehouse password* is to *call any word to mind* if you have committed, ahead of time, to making whatever word you call to mind the treehouse password. All it takes to *choose a place to meet Ben* is to *choose a restaurant* and think that the restaurant is the place to meet Ben. That can all be part of your practical conception of what you are doing in choosing a place to meet Ben. These examples demonstrate that it is possible for (iii) to hold while (i) and (ii) do.

Note that it's possible to meet (iii) even when Ψ-ing as such is factive. This is the case with the restaurant math example. All it takes to *determine the total cost of the meal*—that is, to make a true judgment of that cost – is to *add 25 and 34* and to think of that action in a particular way while you do so. That's because *adding 25 and 34* is also factive.

However, it is not true for any pair of contentful mental action types Φ and Ψ that all it takes to Ψ is to Φ while thinking of your Φ-ing in a particular way. For example: even if you think of *imagining a blue unicorn* as a way of *finding a cure for cancer*, and you act on an intention to *imagine a blue unicorn in order to find a cure for cancer*, it is not sufficient in these circumstances to *imagine a blue unicorn*, and think of it in all the implicated ways, in order to actually *find a cure for cancer*. In short, condition (iii) is placing a substantive constraint here.

Now, given just (i) is it possible for (ii) to be met? It certainly seems so. Condition (i) stipulates that you think of Φ as a way to Ψ. If you think that, it should certainly be possible to form and act on an intention to Φ in order to Ψ, led by this conception. There is nothing particularly remarkable about acting on a complex intention to *Φ in order to Ψ* when you think that Φ-ing is a way of Ψ-ing in your circumstances. This is common sense.

All that remains to do now is to show that it is possible to meet condition (i). Is it possible to think of Φ-ing as a way to Ψ in your circumstances, because Φ-ing bears some particular relation *r* to Ψ-ing in your circumstances?

Given the examples discussed above, it might seem obvious that this is possible. You can certainly use a relation between Φ-ing and Ψ-ing in your circumstances to come to think that Φ-ing is a way to Ψ in your circumstances. The examples of content plurality in mental action given above often recruit contingent features of your environment that rationalize connecting these actions in this way. The sum of 25 and 34 just was the total cost of your meal in dollars, and you get to pick the treehouse password, so any word you consider is one you can pick.

On the other hand, it might seem as though there is an additional problem to be solved here. It can seem as though something spooky is going on, because thinking of Φ-ing as a way of Ψ-ing here is part of what makes your eventual Φ-ing also constitute a Ψ-ing. For your *adding 25 and 34* to constitute an action of *determining the total cost of the meal*, you needed to think doing the former was a way of doing the latter. In order for your *calling a word to mind* to also be a *choice of a treehouse password*, you needed to think doing the former was a way of doing the latter. Now that we've seen the connection between the way you think of your action, and what your action becomes, meeting condition (i) can seem mysterious. Meeting (i) seems to put into play an odd self-fulfilling prophecy. How can you think of Φ-ing as a way of Ψ-ing, if thinking of Φ-ing as a way of Ψ-ing is part of what is needed to make your ultimate Φ-ing into a Ψ-ing after all?

To see that it is not impossible to meet (i), let's take a look at a broader pattern in the philosophy of action. Your practical conception of what you are doing does affect what it is that you actually do, in many circumstances. In many cases of complex action, your thinking of what you are doing as Φ-ing in order to Ψ is necessary for making your Φ-ing constitute a Ψ-ing at all. But that does not keep you from seeing Φ-ing as a way Ψ-ing from the outset.

Let's return to another past example. You can think of *knitting a scarf* as a way of *making your brother a gift* even though your *making your brother a gift* just in *knitting a scarf* depends on your already thinking of your action in that way. When you are acting on the intention to *knit a scarf in order to make your brother a gift*, you have in mind a practical conception of what you are doing. This practical conception includes a relation between *knitting a scarf* in this context and *making your brother a gift*. Simply stated, this practical conception just involves a commitment to the scarf that is to be knit *as* a gift for your brother. You wouldn't be *knitting a scarf in order to make a gift for your brother* without having this relation in mind.

Now consider what happens when you finish knitting the scarf. You already think of the scarf as a gift for your brother; that much is contained in your practical conception of what you are doing. But then to finish the scarf is to finish an item you are already committed to giving to your brother. Since all it takes to make this scarf a gift is to think of it in this way, you have finished making the gift for your brother just by finishing knitting the scarf. (As above: gift wrap and a card are non-essential.)

This example should help us see that there is no additional puzzle to be solved about how you could see *knitting a scarf* as a way of *making a gift for your brother* from the outset. You have in mind a local relation that connects these two actions: you think of a scarf (yet to be made) as a gift (yet to be made). What's more, the fact that one action can execute

both intentions—to *knit a scarf* and to *make a gift for your brother*—only when you intend them all at once, in this structured way, does not keep you from having and acting on some such structured intention. On the contrary, the fact that doing all that at once is within your control partly explains your ability to form this intention in the first place.

The more general phenomenon in action theory is one in which your practical conception of what you are doing – e.g., as *Φ-ing in order to Ψ* – affects the nature of what you actually do by Φ-ing. It may well be that having that practical conception of what you are doing, as a result of acting on an intention to *Φ in order to Ψ*, is required for your Φ-ing to constitute a Ψ-ing as well. But that does not preclude you from having the antecedent belief that Φ-ing is a way of Ψ-ing in your circumstances. On the contrary, it gives that antecedent belief its warrant: it is within your control to Ψ just in Φ-ing, by thinking of your action in that way. This is what makes sense of your antecedent belief that Φ-ing is indeed a way of Ψ-ing.

This structure is ubiquitous in complex intentional action. It is familiar in the philosophy of action from Aquinas's point, famously quoted by Anscombe, that practical knowledge is the *cause of what it understands* – that is, a formal, not (only) efficient, cause of that.[14] To adapt this point to our own discussion: a complex practical conception partly makes your means constitute your complex action as the kind of intentional action that it is.

Let's pause and take stock. In the first part of this section, I presented sufficient conditions on a mental action's having content plurality. In the next part, I explained how it was possible to meet these conditions together. First, I showed that they are the sufficient conditions they presume to be. Then I showed it was possible: to meet (v) when (i) – (iv) hold; to meet (iv) when (i) – (iii) hold; to meet (iii) when (i) and (ii) hold; to meet (ii) when (i) holds; and, finally, to meet (i). In sum, this demonstrates that it is possible to perform a mental action that has one content by performing another mental action that has a qualitatively distinct content. That is, it's possible to perform a complex mental action with content plurality.

## 2.4 The philosophical importance of content plurality

Mental actions can have content plurality in part because an agent's Φ-ing can also constitute a Ψ-ing. A constitution relation between a token Φ-ing and a token Ψ-ing does not imply that all Φ-ings constitute Ψ-ings, and it does not require the agent of a mental action with content plurality to accept any more general constitution relationship between Φ-ings and Ψ-ings outside of her context.

This is why the model of mental actions with content plurality I have developed here is a powerful tool of philosophical explanation. It allows

us to thread the needle between (a) implausible type-level constitution claims of two distinct action types and (b) implausible dissociation between two token actions which need to be more closely related for some local explanatory purpose. Threading this needle allows us to sew up several holes in distinct philosophical debates.

In this section, I'll apply this tool to three philosophical questions: a question about transparent self-knowledge of belief and intention; a question about the relation between theoretical judgments and practical decisions; and a question about the nature of inference. In each case, we will see that a token contentful mental action's constituting another token mental action with distinct content without a type-level constitution relation can help clear up philosophical confusion.

### *2.4.1 Transparent self-knowledge*

You can come to know whether you believe that *p* in part by figuring out whether *p* is true. Similarly, you can come to know whether you intend to Φ by determining whether to Φ. Self-attributions of beliefs and intentions are thus thought to be 'transparent' in a particular sense: the question of what you believe is 'transparent' to the question of what is true, and the question of what you intend is 'transparent' to the question of what to do.[15]

Importantly, this transparency of one question to another is thought to be a special feature of the first-personal perspective. Epistemologists working on this topic often take it that explaining how you can transparently self-attribute beliefs and intentions will illuminate the other special features of the first-personal perspective each of us has on her own beliefs and intentions—including the authority and epistemic privilege of this position.

Why does it make sense, from the first-personal perspective, to trade questions in this remarkable way? In the belief case, we can ask: how could my judgment that *p* contribute to my judging that *I believe that p*? In the intention case, we can ask: how could a decision to *Φ* contribute to my judging that *I intend to Φ*? In neither case does there seem to be a justificatory relationship between the contents of the relevant mental actions. Thus, it doesn't help to interpret transparent self-attribution as inferential in either the belief or the intention case.[16]

It does help to see transparent self-attributions of beliefs and intentions as mental actions with content plurality. This shifts the burden of explanation away from a general relationship of inferential support between contents and towards a local understanding of how an agent could do two particular things at once in thought. In the belief case, we can come to see how an agent could judge that *p* and judge that she believes that *p* all at once. In the intention case, we can come to see how an agent could decide to Φ and judge that she intends to Φ all at once.

Elsewhere I have given a sustained explanation of transparent self-knowledge of belief.[17] Here I will give the explanation of transparent self-knowledge of intention.

Let's see how conditions (i) – (v) can be met in transparent self-attribution of intention. You can make a decision and make a judgment about what you intend to do all at once when:

i you think that *deciding whether to Φ* is a way to *figure out whether you intend to Φ* in your circumstances because you intend to do what you decide to do;
ii you act on an intention to *decide whether to Φ in order to figure out whether you intend to Φ*, led by the conception of these actions mentioned in (i);
iii all it takes in your circumstances to figure out whether you intend to Φ is to think of the content of a token decision whether to Φ as what you intend to do; and
iv you execute both intentions (to *decide whether to Φ* and to *figure out whether you intend to Φ*) just by *deciding whether to Φ* intentionally in such a way that
v the content of your *deciding whether to Φ* as such is qualitatively distinct from the content of your *figuring out of whether you intend to Φ*.

Let's take a case in which you do decide to Φ, and in so doing you figure out that you intend to Φ. In this case, there is one token mental action that both executes the intention to *decide whether to Φ* and the intention to *figure out whether you intend to Φ*. It executes the former as a decision to Φ. It executes the latter by also constituting a judgment that *you intend to Φ*. In this case, your *figuring out whether you intend to Φ* has content plurality.

What relation *r* between *deciding whether to Φ* and *figuring out whether you intend to Φ* informs your mental activity in this case? This relation, or rather your understanding of it, is crucial to the content plurality of your ultimate mental action. This relation makes it into your practical conception of what you are doing and makes it the case that your *decision to Φ* is also a *judgment that you intend to Φ*. Why does it make sense to structure your activity in this way?

The relationship is one between what you decide and what you intend. Insofar as you understand that what you decide to do just is what you intend to do (at least for this very moment), you can use this procedure to self-attribute intentions.

So transparent self-attribution of intention involves performing a mental action with content plurality. Why does this help us understand the epistemology of this self-knowledge?

This model of transparent self-attribution provides a strong explanation of the epistemic credentials of such attributions. Any self-attribution

of intention made in this way will be true, as making a decision to Φ is sufficient for (contemporaneously) intending to Φ.[18] The self-attribution is warranted as a self-attribution of an intention because you know what kind of thing you are doing in thought – namely, *deciding whether to Φ* – and that doing that sort of thing is sufficient for figuring out what you intend to do.[19] It is warranted as an attribution of intention to yourself in particular because this procedure requires no identification of yourself among others; it is thus immune to error through misidentification.[20] Finally, it is warranted as a self-attribution of an intention to *Φ*, in particular, by the consciousness of the content of your decision to *Φ*.[21]

There are two key advantages of this account of transparent self-knowledge of intention.

First, the account specifies a plausible relation between a token decision to *Φ* and a judgment that *you intend to Φ*. The former can constitute the latter. The account explains how that can be the case even though not all decisions constitute judgments about what you intend.

Second, it allows for a remarkably secure form of warrant in at least one respect. Since the self-attribution of an intention is made at precisely the same moment as a decision to *Φ*—by way of this one token mental action—and a decision to *Φ* is sufficient for at least contemporaneous temporary intention to *Φ*, this kind of self-attribution of intention cannot err by way of a temporal mismatch between decision and self-attribution.[22]

There are certainly more questions to be answered about this account of transparent self-attribution of intention. I set these aside to save space. What is most important to see here is just that content plurality in mental action transforms our inquiry into such self-attribution. Instead of asking about the relationship between two temporally distinct mental actions, this model of content plurality in mental action allows us to consider another strong explanation of the epistemology of transparent self-attributions.

### *2.4.2 Judgments and decisions*

This understanding of mental actions with content plurality also generates a subtle new interpretation of practical reasoning that involves both judgments and decisions.

There is a longstanding debate about how your judgments with contents of the form *I ought to Φ* or *Φ-ing is the thing to do* relate to decisions to *Φ*. At first, these sorts of mental events seem to belong to two different categories, the 'theoretical' and the 'practical': any judgment is a doxastic acceptance of a truth-evaluable content, whereas a decision *to Φ* is directed at action, and thus takes a fundamentally action-guiding attitude towards its content. On further reflection, though, some particular 'theoretical' judgments seem to have immediate and not merely

causal impact on my plans, or even on my actions themselves.[23] Settling what I ought to do, or settling what the thing to do is, can simply settle for me what I'll do.

Partly to close the uncomfortable gap between a judgment that *Φ-ing is the thing to do* and a decision to *Φ*, Alan Gibbard has influentially argued that making that sort of judgment just is deciding to *Φ*.[24] Proposing this perfectly general connection between judgments of this form and decisions allows Gibbard to develop a systematic explanation of practical reasoning, modeled on certain central key features of 'theoretical' reasoning.

But Gibbard's proposal faces a simple problem of extensional inadequacy. It is usually but not always the case that a judgment with content like *Φ-ing is the thing to do* counts as a decision to *Φ*. One way to see this is to consider certain cases of 'weakness of the will' in which an agent takes some action Φ to be unequivocally *the thing to do*, but does not thereby decide to *Φ*.[25] This kind of situation can seem painfully familiar. And if it is so much as possible, then Gibbardian expressivism must fail in its full generality.

A better picture of the relationship between such judgments and decisions to act would allow for certain judgments to constitute decisions without implying that all judgments constitute such decisions. This is precisely what content plurality in mental action allows. Some judgments that *Φ-ing is the thing to do* also constitute decisions to *Φ*, but some do not.

Let's consider how a judgment that *Φ-ing is the thing to do* can also constitute a decision to *Φ*. In some cases, you are already committed—in the distinctively practical sense in which decisions and intentions are commitments—to doing whatever is *the thing to do*. If so, you can use this practical commitment to structure your thought. That is, you can see *figuring out the thing to do* as a way to *decide what to do*. Acting on an intention to *figure out the thing to do in order to decide what to do* will set up a context in which one and the same mental action can execute both these intentions at once. Consider a case where

i you think that *figuring out the thing to do* is a way to *decide what to do* in your circumstances because you commit to doing *the thing to do* (de dicto);
ii you act on an intention to *figure out the thing to do in order to decide what to do*, led by the conception mentioned in (i);
iii all it takes to decide what to do in your circumstances is to commit to doing that which you figure out is the thing to do; and
iv you execute both intentions (to *figure out the thing to do* and to *decide what to do*) just by *figuring out the thing to do* intentionally in such a way that
v the content of your *figuring out the thing to do* is qualitatively distinct from the content of your *deciding what to do*.

In this case, your decision has content plurality. It has the content of a decision to Φ (say), but it is constituted by a judgment with the distinct content that *Φ-ing is the thing to do*.

In cases like these, there is a tight, non-accidental constitution relation between such judgments and their related decisions. However, it is not a general connection that holds on the level of types of mental actions. Understanding this connection in some circumstances does not require us to say that any specific kinds of judgments necessarily, or even universally, constitute decisions to act as well. It does not require any significant revision of the typology of mental actions. For that reason, allowing cases like this does not implausibly rule out certain kinds of weakness-of-will cases, in which you fail to decide to do that which you think is the thing to do.

Content plurality in mental action also lets us see how judgments of other forms can constitute decisions in certain contexts. You might, on Saturdays, be committed to doing whatever is most fun. In this case, you might see *figuring out what is most fun to do* as a way to *decide what to do*. If you then act on an intention to *figure out what is most fun to do in order to decide what to do*, and you execute the relevant intentions in one mental action, that action will be a judgment with the content *Φ-ing is most fun to do* that also constitutes a decision to *Φ*.

I do not mean to imply that the explanatory power of content plurality in mental action rivals the full power of Gibbardian expressivism. It should nonetheless be an attraction of the picture I have developed here that it allows us to see how certain judgments are also decisions without committing to a full expressivist picture with all its (apparently) revisionary implications. Content plurality reshapes the debate about judgments and decisions by opening up new view on which certain token judgments constitute token decisions although no such constitution relation holds at the type level.

### *2.4.3 Inference*

The structure of content plurality in mental action also helps solve a problem in another philosophical debate: the debate about the nature of inference.

It is usually thought that inferring—e.g., inferring *q* from *p*—is a form of mental action. It is also usually thought that inference from *p* to *q* involves a movement in thought from a judgment that *p* to a distinct judgment that *q*.[26] This second assumption implies that a kind of transition between distinct mental actions constitutes inference.

This gives rise to recalcitrant problems for theories of inference. There are many ways of moving from one judgment to another in thought, not all of which constitute inferential transitions. To make an inference between two distinct judgments, the first judgment must be one

in which the first judgment is taken to warrant the second. As Gottlob Frege famously put it, "to make a judgment because we are cognisant of other truths as providing a justification for it is known as inferring."[27] Paul Boghossian has influentially framed this constraint as the "(Taking Condition): Inferring necessarily involves the thinker taking his premises to support his conclusion and drawing his conclusion because of that fact."[28] The next task is to say what this 'taking' comes to, and how it enters into any inference.

Not everyone accepts the Taking Condition, in part because its acceptance seems to lead to regress or deviant causal chains in an account of inference.[29] I'll briefly summarize why.

We can understand the inferrer's taking as either an occurrent mental event—like a judgment—or as a belief, as a standing state. If the inferrer's taking some premise to support his conclusion is a judgment, then we can ask how it modulates a transition from a judgment that *p* to a judgment that *q* to make the transition itself into an inferential one.

Lewis Carroll's famous parable of the tortoise and Achilles has clarified that this taking judgment could not act as just another premise judgment in the course of the transition, on pain of regress.[30] If you needed to judge *if p then q* between a judgment that *p* and a judgment that *q* in order to make the whole move a transition, then surely you would also need to judge *if p and (if p then q), then q* somewhere in the middle to make the transitions among the three judgments a form of inference. And so on.

We might try to bridge the gap between a judgment that *p* and a judgment that *q* by suggesting that the judgment that *p* needs, additionally, to cause the judgment that *q*. But there are many deviant—and thus non-inferential—ways for a judgment that *p* to cause a judgment that *q*. A judgment that *I'm about to fall off this cliff* might cause me to faint, and my fainting might cause me to judge (after I wake up) that *I have fainted*. This is a case in which, by transitivity of causation, my judgment that *I'm about to fall off this cliff* causes (downstream) a judgment that *I have fainted*, but it is certainly not an inference from one to the other.

We might instead interpret the inferrer's taking the first judgment to support the second one as a standing belief about the justificatory relationship between the contents *p* and *q*. But you can have this standing belief without using it in a transition from a judgment that *p* to a judgment that *q*. In that case, the transition would still not count as an inference.

There are other ways to understand the taking involved in the Taking Condition.[31] But none seems particularly well suited to explain how taking ensures an inferential transition.

This leaves us at an impasse. Some have tried to escape it by rejecting the Taking Condition. But the main problem for theories of inference is not the Taking Condition. Theories of inference have been significantly weakened

by the assumption that any inference must be executed in transition from one judgment to another. The availability of content plurality in mental action can help us see why this assumption is not necessary. That is because you can make an inference all at once. One token mental action can be a judgment that *p* that also constitutes a judgment (therefore) that *q*.

The easiest case to understand is a case in which you take there to be a biconditional relationship between *p* and *q*. If you believe that *p iff q*, then you can see that *figuring out whether p* is a way of *figuring out whether q*. Here's how conditions (i) – (v) are met here:

i you think that *figuring out whether p* is a way to *figure out whether q* in your circumstances because *p* has the same truth value as *q*;
ii you act on an intention to *figure out whether p in order to figure out whether q*;
iii all it takes to figure out whether *q* in your circumstances is to think of the truth value of the content of a token judgment whether *p* as the truth value of *q*; and
iv you execute both intentions (to *figure out whether p* and to *figure out whether q*) just by intentionally *figuring out whether p* in such a way that
v the content of your *figuring out whether p* is qualitatively distinct from the content of your *figuring out whether q*.

In this case, let's say you execute both intentions in a judgment that *p* which also constitutes a judgment that *q*. As the former, it executes the intention to *figure out whether p* (assuming, as we can for the time being, that *p* is true). Since it constitutes the latter, it also executes the intention to *figure out whether q*. You can do all this at once because acting on an intention to *figure out whether p in order to figure out whether q* involves having in mind a practical conception of what you are doing that incorporates an understanding that *p* has the same truth value as *q*. Your having this practical conception in mind makes your judgment that *p* into a judgment that *q* as well.

It is important to see that there is no room for deviant causation here. Your taking *p* to support *q* is not just one cause among many, nor just one judgment between others, nor a background commitment that may or may not get used in occurrent thought. Your taking *p* to support *q* makes your judgment that *p* also a judgment that *q* in this particular context.

Might there nonetheless be deviance in the way that you come to execute your several intentions? Not while we understand an execution of an intention in the way I have defined it. You execute an intention to *Φ* just when you do in fact Φ intentionally in acting on this very intention to *Φ*. What's more, Φ-ing intentionally rules out deviance.[32]

Just recognizing the possibility of making an inference all at once may not yet give us a complete theory of inference. But it rejects a received

assumption in the philosophical discussion about inference and thus significantly reshapes the discussion.

## 2.5 Objections

Here I will address two objections to content plurality in mental action.

### *2.5.1 Overcrowding*

You might feel that mental actions with content plurality are simply too crowded.[33] You can't fit all that content in there. This objection is more intuitive than precise, but it's worth addressing it in order to see how it stems from a misunderstanding.

By explaining how one mental action can execute more than one intention at once, in part by constituting an action of another kind as well, I mean to have located room for mental actions to have more than one content at once. I have used no ad hoc stipulations to motivate this model; I have instead relied on more general claims about action in order to explain how content plurality in mental action is possible.

It is also important to note that I do not mean to suggest that mental actions can have endless content plurality. There are limits on the total number of things you can do at once in thought, and these limits may derive primarily from limits on working memory. Even though it is possible to, say, *think of a sentence in order to think of something Harry might say to Ron*, it does not seem possible to *think of a sentence of English that is easily translatable into Russian in order to think of something Harry might say to Ron in order to tell my editor how to replace the twelfth sentence on p. 97…* all at once. Even though there might, in principle, be one mental action that executes all the many intentions implicated in this multiply complex intention, it does not seem possible for an agent to act on all those intentions at once, while having in mind the practical conception that doing all that at once would require.

However, the impossibility of acting on such intentions does not imply the impossibility of acting on smaller combinations. If we restrict our consideration to a reasonable amount of content plurality, we have no need to feel spooked by its very possibility.

This objection does not seriously threaten the proposal developed here. But I will venture a diagnosis of its source nonetheless. It is easy to slip into thinking of mental actions – and indeed all thoughts, active or passive – as 'internal' utterances of sentences. Because it's hard to see how a *sentence* could have content plurality, this understanding of thought might make mental actions with content plurality look particularly crowded. But there are already good independent reasons not to think of thoughts as 'internal' utterances of sentences. One of the most important reasons not to do this has to do with the attitudinal aspects

of thought. For example: if a judgment could only consist in an 'internal' utterance of a sentence, it is not clear why judging that *p* would ever constitute the doxastic commitment that judging that *p* really does constitute. Utterances can always be sincere or insincere, but there is no such thing as a sincere or insincere judgment. Recognizing the normative commitments involved in occurrent thoughts already recommends against any model on which thoughts are 'internal' utterances at all.

### *2.5.2 Complex contents*

Another way to resist the content plurality of mental actions is to insist that each putative example of content plurality is really an example of content unity, where the content in question is logically complex. A mental action that I identify as a judgment that *p* which constitutes a judgment that *q* might be re-interpreted as a judgment that *p and q*, and so on.

There is a simple reason that this re-interpretation cannot work in full generality. Some cases of content plurality are cases in which one action with a certain content is constituted by an action with another kind of content—one with which its own content cannot be logically concatenated. To recall an example from above: the content of a *calling a word to mind* is a linguistic item—here, a word—and the content of a *decision on a treehouse password* is plausibly propositional. These contents cannot be conjoined into a meaningful complex content. At least we cannot use truth-functional connectives to concatenate a word and a proposition in such a way as to get just one logically complex content.

In some cases of mental action with content plurality, this re-interpretation looks better. In particular, it looks better when one judgment constitutes another. Even in these cases, though, the reinterpretation would eliminate the explanatory power of the tool that emerges from the model of content plurality I have developed here.

Consider transparent self-attribution of belief. It would not help us see how you can judge *I believe that p* by judging *p* if we thought that a logically complex content—*p and I believe that p*—is judged in transparent self-attribution of belief. Similarly, we could not use the structure of intentional action to solve any deviance problems for inference if an inference from *p* to *q* executed all at once simply had the content *p and q*. Instead, the etiology of that logically complex judgment itself would have to be explained—perhaps by yet another inference.

## 2.6 Conclusion

There are mental actions with content plurality. These are complex mental actions with certain contents that are constituted by mental actions with qualitatively distinct contents.

For two distinct types of contentful mental action Φ and Ψ, if

i you think that Φ-ing is a way to Ψ in your circumstances, because Φ-ing bears a certain relation *r* to Ψ-ing in your circumstances;
ii you act on an intention to *Φ in order to Ψ*, led by this conception of Φ-ing;
iii all it takes to Ψ in your circumstances is to think of a token Φ-ing of yours as bearing that same relation *r* to Ψ-ing, and
iv you execute both intentions (to *Φ* and to *Ψ*) just by Φ-ing intentionally in such a way that
v the content of your Φ-ing is qualitatively distinct from the content of your Ψ-ing,

then your Ψ-ing has content plurality.

A mental action of Φ-ing can also constitute a Ψ-ing because you are thinking of what you are doing in a certain way. Your practical conception can make your Φ-ing constitute a Ψ-ing too.

On this model of content plurality in mental action, a token Φ-ing can constitute a token Ψ-ing even though there is no more general constitution relationship at the type level. This local token constitution without implausible type constitution allows us to use content plurality as a tool of philosophical explanation in several distinct debates. Content plurality in mental action helps to explain transparent self-knowledge of intention and of belief. It gives us a new way of understanding the relationship between 'theoretical' judgments and practical decisions. And it urges us away from a traditional assumption that an inference must be a transition between temporally separate judgments.

There is no general way to understand mental actions with content plurality as single actions with single, logically complex contents. But this does not mean that mental actions with content plurality are illicitly overcrowded in any way. In order to see how a mental action can have content plurality, we need to move away from the independently problematic, though natural, understanding of thoughts as internal utterances.

## Notes

1. For extensive feedback, I'm grateful to Michael Brent, Christopher Peacocke, and an anonymous reviewer for this volume. I'm also indebted to audiences at the University of Warwick, CUNY Graduate Center, the City College of New York, and New York University. Thanks, everyone.
2. Words in bold indicate technical terms to be defined.
3. Small capital letters denote concepts. Note that judgment is the mental event with propositional content that is normatively and descriptively governed by truth. I borrow this formulation from Shah and Velleman (2005) and apply it here to individuate judgments among mental events.
4. For more on non-voluntarism, see Shah and Velleman (2005), Williams (1976) and A. Peacocke (2017).

5. Italics throughout do not indicate emphasis. Rather, italics identify intensional descriptions of actions that capture the contents of an agent's thoughts or intentions.
6. You can, of course, do all sorts of other things in order to indirectly bring it about that you do not think about a polar bear. For more on this kind of extended mental action, see Mele (2009).
7. Compare Anscombe (1957), Davidson (1967/2001, 1968/2001). It is better to say that actions are intentional under aspects, not descriptions, as there is nothing essentially linguistic about the understanding you must have here. Nonetheless, I follow convention in using the term "description."
8. Anscombe (1957, p. 34ff.) and Davidson (1963/2001, 1971/2001, 1978/2001). They both put the point in terms of one action's being intentional under several descriptions at once. I favor instead the metaphysics of constitution here: one action intentional under one specific description can constitute an action intentional under another description.
9. Action variables ($\Phi$, $\Psi$) used in these italicized intensional contexts do not refer to letters, but to the actions to which they refer. The italicized portions of this chapter are thus a form of Quine quotation.
10. I do not yet want to claim that these jointly sufficient conditions are also individually necessary conditions on a mental action's having content plurality, although I don't want to rule this out, either.
11. Anscombe (1957, p. 67ff); cf. Hampshire (1959). Anscombe defines "practical knowledge" in two ways, as knowledge of what happens in a particular instance of action and then later as a kind of capacity to do something, like know-how. The connection between these two is a matter of some discussion; see, e.g., Frost (2019). It's only the first of these two senses of "practical knowledge" that I mean to target here, with no commitment concerning the relationship between the two senses. On practical knowledge as knowledge, see Ford, Hornsby, and Stoutland (2014), Haddock (2014), Moran (2004), Schwenkler (2012, 2015), and Velleman (2007).
12. Cf. Schwenkler (2015).
13. Pace Davidson's (1978/2001) discussion of the carbon copies. See Setiya (2008) for a discussion. Even if practical conception did not necessarily involve doxastic commitment, there would be cases in which doxastic commitment is involved in practical conceptions, and some of the mental actions in question would be examples. This would be enough for our purposes.
14. Anscombe (1957). Compare Schwenkler (2015) on the role of formal cause in this discussion. Setiya (2016a) powerfully argues that this is a restricted claim in Anscombe. The restriction he endorses does allow application to the cases of complex intentional action considered here.
15. The most famous formulation of transparency of belief is found in Evans (1982), although as Moran (2001) notes, a clear formulation was also in Edgley (1969). For discussions of transparent self-knowledge, see also Barnett (2015), Byrne (2011), Cassam (2014), Paul (2012), and Way (2007).
16. See Byrne (2011) for the view that transparent self-attribution is inferential, and see Barnett (2015) and my (2017) for criticism of this view.
17. A. Peacocke (2017).
18. But see Paul (2012) for an argument that decision isn't sufficient for even contemporaneous intention.
19. Here it matters that a practical conception can indeed amount to knowledge of what you are doing. Practical conceptions that amount to knowledge do so because the agent is acting in a way she controls. So, it also

matters here that you control the kind of thing you are doing in thought. I think this control condition is fulfilled in this context. For more on the connection between control and practical knowledge, see e.g., Velleman (2007).

20. Pryor (1999), Shoemaker (1968).
21. These explanatory features exactly mirror those presented in A. Peacocke (2017). Further (analogous) explanation of each point can be found there.
22. This last point suggests a further application of content plurality in mental action as well. Content plurality in mental action can be used to formulate a new model of Descartes's (1988) *cogito*. You can, at any point, perform *any* mental action in order to ensure certainty of your existence. One and the same mental action can be both a *Φ*-ing (for any mental action type Φ) and a judgment that you exist.
23. See the debate between internalism and externalism about practical reason, e.g., Wallace (2001, 2014).
24. See Gibbard (2003). It is key that Gibbardian expressivism is in the first instance a theory about *judgments* – and only derivatively about speech acts like assertions (p.76).
25. I do not here mean to imply that all cases of weakness of the will are structured in this way. Some cases of weakness of the will simply involve failure to act on a pre-existing decision to act.
26. See Boghossian (2014), Broome (2014-comment), Wedgwood (2006), Wright (2014-comment), McHugh and Way (2016, 2018). Neta (2013) is a refreshing exception.
27. Frege (1979, p. 3), quoted in Boghossian (2014, p. 4).
28. Boghossian (2014, p. 5).
29. See, e.g., McHugh and Way (2016) and Siegel (2019).
30. Carroll (1895).
31. See Sections 7 ("An intuitional construal of taking") and 9–12 (on rule-following) of Boghossian (2014).
32. This creates its own deviance problem for definitions of action, as Davidson (1971/2001) discussed.
33. I'm grateful to Brie Gertler for raising a version of this thought in personal correspondence.

## References

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

Barnett, D. J. (2015). Inferential justification and the transparency of belief. *Noûs*, *50*(1), 1–29.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*, 1–18.

Broome, J. (2014a). Normativity in reasoning. *Pacific Philosophical Quarterly*, *95*, 622–633.

Broome, J. (2014b). Comments on Boghossian. *Philosophical Studies*, *169*, 19–25.

Byrne, A. (2011). Self-knowledge and transparency I: Transparency, belief, intention. *Proceedings of the Aristotelian Society*, *LXXXV*, 201–221.

Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, *4*(14), 278–280.

Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.

Davidson, D. (1963/2001). Actions, reasons, and causes. In *Essays on actions and events* (pp. 3–20). Oxford: Clarendon Press.

Davidson, D. (1967/2001). The logical form of action sentences. In *Essays on actions and events* (pp. 105–121). Oxford: Clarendon Press.

Davidson, D. (1971/2001). Agency. In *Essays on actions and events* (pp. 43–62). Oxford: Clarendon Press.

Davidson, D. (1978/2001). Intending. In *Essays on actions and events* (pp. 83–101). Oxford: Clarendon Press.

Descartes, R., Cottingham, J., Stoothof, R., & Murdoch, D. (Trans.). (1985). *Selected philosophical writings*. Cambridge: Cambridge University Press.

Edgley, R. (1969). *Reason in theory and practice*. London: Hutchinson.

Evans, G. (1982). *The varieties of reference*. New York: Oxford University Press.

Ford, A., Hornsby, J., & Stoutland, F. (Eds.). (2014). *Essays on Anscombe's intention*. Cambridge, MA: Harvard University Press.

Frege, G. (1979). Logic. In H. Hermes, F. Kambartel, & F. Kaulbach (Eds.), & Long, P., & White, R., (Trans.). *Posthumous writings* (pp. 1–8). Oxford: Basil Blackwell.

Frost, K. (2019). A metaphysics for practical knowledge. *Canadian Journal of Philosophy*, *49*(3), 314–340.

Gibbard, A. (2003). *Thinking how to live*. Cambridge, MA: Harvard University Press.

Haddock, A. (2014). The knowledge that a man has of his intentional actions. In A. Ford, J. Hornsby, & F. Stoutland (Eds.), *Essays on Anscombe's intention* (pp. 147–169). Cambridge, MA: Harvard University Press.

Hampshire, S. (1959). *Thought and action*. Notre Dame, IN: University of Notre Dame Press.

McHugh, C., & Way, J. (2016). Against the taking condition. *Philosophical Issues*, *26*(1), 314–331.

McHugh, C., & Way, J. (2018). *What is reasoning? Mind*, *127*(505), 167–196.

Mele, A. (2009). Mental action: A case study. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 17–37). New York: Oxford University Press.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton, NJ: Princeton University Press.

Moran, R. (2004). Anscombe on 'practical knowledge.' In J. Hyman, & H. Steward (Eds.), *Royal institute of philosophy supplement* (pp. 43–68). Cambridge: Cambridge University Press.

Neta, R. (2013). What is an inference? *Philosophical Issues*, *23*(1), 388–407.

O'Brien, L., & Soteriou, M. (Eds.). (2009). *Mental actions*. New York: Oxford University Press.

Paul, S. K. (2012). How we know what we intend. *Philosophical Studies*, *161*, 327–346.

Peacocke, A. (2017). Embedded mental action in self-attribution of belief. *Philosophical Studies*, *174*(2), 353–377.

Peacocke, C. (2008). *Truly understood*. New York: Oxford University Press.

Pryor, J. (1999). Immunity to error through misidentification. *Philosophical Topics*, *26*(1), 271–304.

Ryle, G. (1971a). A puzzling element in the notion of thinking. In *Collected papers, volume II: Collected essays 1929–1968* (pp. 391–406). London: Hutchinson.

Ryle, G. (1971b). Thinking and reflecting. In *Collected papers, volume II: Collected essays 1929–1968* (pp. 465–479). London: Hutchinson.

Ryle, G. (1971c). The thinking of thoughts: What is 'Le Penseur' doing? In *Collected papers, volume II: Collected essays 1929–1968* (pp. 480–496). London: Hutchinson.

Schwenkler, J. (2015). Understanding practical knowledge. *Philosopher's Imprint, 15*(15).
Schwenkler, J. (2012). Non-observational knowledge of action. *Philosophy Compass*, *7*(10), 731–740.
Setiya, K. (2008). Practical knowledge. *Ethics*, *118*, 388–409.
Setiya, K. (2016a). Anscombe on practical knowledge. In *Practical knowledge: Selected essays* (pp. 156–170). New York: Oxford University Press.
Setiya, K. (2016b). *Practical knowledge: Selected essays*. New York: Oxford University Press.
Shah, N., & Velleman, D. (2005). Doxastic deliberation. *The Philosophical Review*, *114*(4), 497–534.
Shoemaker, S. (1968). Self-reference and self-awareness. *The Journal of Philosophy, LXV*, 556–579.
Siegel, S. (2019). Inference without reckoning. In B. Balcerak Jackson, & M. Balcerak Jackson (Eds.), *Reasoning: New essays on theoretical and practical thinking* (pp. 15–31). New York: Oxford University Press.
Smith, C. S. (1991). *The parameter of aspect*. Dordrecht: Springer Science + Business Media.
Soteriou, M. (2013). *The mind's construction: The ontology of mind and mental action*. New York: Oxford University Press.
Velleman, D. (2007). What good is a will? In A. Leist (Ed.), *Action in context* (pp. 193–215). Berlin: Walter de Gruyter.
Wallace, R. J. (2001). Normativity, commitment, and instrumental reason. *Philosopher's Imprint*, *1*(3), 1–26.
Wallace, R. J. (2014). Practical reason. In E.N. Zalta (Ed.). Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/practical-reason/, October 28, 2019.
Way, J. (2007). Self-knowledge and the limits of transparency. *Analysis*, *67*(3), 223–230.
Wedgwood, R. (2006). The normative force of reasoning. *Noûs*, *40*(4), 660–686.

# 3 Attending as Mental Action

## Mental Action as Attending[1]

*Wayne Wu*

The near palindrome in the title gestures at the thesis: attending is the basic movement of the mind and every movement of the mind is an attending (see also Levy, 2023, this volume). William James's characterization provides the right starting place for thinking about attention:

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.
>
> (James, 1896, pp. 403–404)

I read talk of *the mind's taking possession of a train of thought* broadly as capturing trains of memories, images, propositions, and so forth. Attention gets a foothold when it embodies a type of mental selectivity that exemplifies the agent's agency. In the case of mental action, this selectivity concerns a traversal across intentional contents.

This paper begins by presenting the framework I shall operate with, namely the idea of actions as occurring in a structured space of behavioral possibilities. I have argued for this picture in various papers, and I shall be brief in presenting the picture. That said, the presentation has not always been perspicuous, so formulations will be tightened in this discussion and crucial clarifications entered. This will allow me to correct certain misinterpretations of the account. I also explicate a notion of an *action capacity* to explain why certain passive behaviors can count as actions. I then explain what it is to simply intentionally attend, focusing on the basic case of covert perceptual attention, the smallest movement of mind. And I then argue that all mental actions are forms of attending.

DOI: 10.4324/9780429022579-4

## 3.1 The many-many problem

If an agent acts, then the agent's acting occurs against the context of a behavior space that presents different action possibilities for a specific agent at a time. This behavior space is constituted (one) by inputs, which are the agent's psychological states at that time such as her seeing, feeling, remembering, entertaining, and so on, and (2) by outputs which can be further psychological states but in general are responses of various sorts. Thus, the responses can involve a movement of the body or a further mental state that is informed by the input. The space maps inputs to outputs, and each input-output path identifies a capacity to act that the agent has at that time, grounded in the input mental state. In sum, the behavior space identifies a complex agentive capacity the subject possesses in that context.

When the agent acts, one of the potential paths is actualized. The agent responds in light of how she *takes things*. Talk of "taking things" is to indicate the subject's intentional orientation as embodied by the input state that comes to guide the action. Accordingly, there can be perceptual, doxastic, mnemonic, as well as imagistic and suppositional takings. We can speak of the subject's taking things to be (the case), takings things for the sake of argument, or simply sensorily taking things in. The appropriate further way of speaking about taking things will be unpacked by the nature of the input psychological state.

The possible actions available to an agent at a given time depends on many things. The potential behavioral paths are fixed by the subject's available agentive capacities as well as the state of her mind such as what she is currently perceiving, thinking, or remembering, by what her internal bodily states are like, say the state or position of her body as well as physiological states, and by what the world is like. Given these dependencies, as the action proceeds, the background behavior space is malleable. At any given moment in time during action, if we were to represent a behavior space at that time, that depiction must be sensitive to the fact that the subject's behavior space will change its structure as new inputs states come on the scene, extant states are altered, and others are lost.

When the agent acts, the framework of the behavior space can capture a dynamic background to the structure of human action. It is important to remember this, for a given behavior space is like a snapshot of the agent's state of mind, but this introduces distortions if we use it to discuss the agent's acting, an extended process. In the latter case, we must remember that the input state plays a dynamic role in the guidance of action, say when a subject's perception of a target continues to inform a movement or stream of thought. The concept of a behavior space provides a theoretical framework to characterize the structure of action as a dynamic phenomenon. The crucial aspects of the picture are the basic ideas of (a) many potential input-output mappings that define (b) the

set of possible behavioral paths that constitute the behavior space at a time, and (c) action as the actualization of one among other paths that involves linking a specific input and a specific output where the former guides the latter.

What does this say about the nature of agency? Not all behaviors are actions, and one behavior that contrasts with action is a reflex. If one were to provide a gloss of this contrast, one might note that action of necessity involves some executive capacity while reflexes are things that we suffer, that happens automatically and to which we are passive. To bring this aspect of reflex out, I define a *pure reflex* as the conceptually relevant contrast, and in the context of behavior spaces, a pure reflex is an output that of necessity is always produced by a specific input. What this means is that for a creature for whom a pure reflex is the only possible behavior, thus a creature that is not an agent in the relevant sense, its behavior space is just a one-one map that holds of necessity. The input is always mapped to the output, and where the input occurs, necessarily so does the output (this sort of reflex is not of the type that we experience, for these can be stopped in some cases or perhaps even fail to be actualized). This eliminates any other possibilities for behavior.

If that much is granted, then a pure reflex, as defined in terms of this necessitated one-one map, is not an action. Conversely, an action cannot be a pure reflex, and this means that the path that constitutes an action cannot be a necessary one. To remove pure reflexes, one must break the necessity between the input and output that defines it. Eliminating necessity entails additional possibility, specifically a branched mapping, say the input being mapped to two outputs or two inputs being mapped to a single output where each link is not necessitated. Since this behavior space identifies more than one behavioral possibility, action requires a branched behavior space. I include here the limiting case of mapping an input to a null response.

As we are interested in human action, a human agent's behavior space typically involves many inputs and many outputs. In other work, I have pointed out that the agent typically must solve a *Many-Many Problem* posed by multiple behavioral options.[2] For given so many options, the agent's acting means that only one of many paths gets actualized. The Many-Many Problem must be solved if there is to be action. There are two versions of the Problem. In my first presentation of my version of selection for action (2008), I noted a *deliberative* and a *non-deliberative* version (see also (Wu, 2011b)). The deliberative version is one whose solution is an intention or similar state. This problem is typically faced in the context of having to choose, a context whose structure is a behavior space that includes a set of possible decisions/choices as output and the agent's take on relevant considerations as input. The subject's practical deliberation proceeds within a conception of relevant parts of that behavior space which are the focus of her deliberation, so a proper

subset of the total possibilities of action available to her at the time of deliberation. The result of deliberation is the formation of an intention to act. The intention counts as a solution to the deliberative version of the Many-Many Problem.

There is a second version of the problem, the *non-deliberative* version, and its solution is the action performed. I noted, in focusing on bodily action, that this version "arises at the beginning of action and extends throughout its course, prompted by the demands of producing and guiding movement" (Wu, 2008, p. 1007). It is at the level of the non-deliberative version that the dynamics of action come to the fore. Intentional action emerges from the convergence of the two problems in that where an intention is the agent's perspective on what is to be done, the intention constituting a solution to the deliberative problem, then the path taken is due to the agent's intending to act in that way; it is an implementation of the solution to the deliberative problem (intention) that amounts to solving the non-deliberative problem because it constitutes action. The action just is the input state's guiding the response.

The a priori argument for a structure for action sets a frame within which we can probe different facets of agency. If that argument is correct, then actions must have a certain structure. My approach then is to understand the *biology* of human agency in light of the necessary framework of the behavior space that we arrived at on a priori grounds. Thus, the generation of action is understood in terms of how an agent's intention biases the agent's orientation toward the world and biases the agent's capacities for response such that the path corresponding to the intended action is taken. Affirming this structural influence of intention is driven by the need to make intelligible why certain paths are taken, namely those paths which constitute the very actions that the agent intends to perform. Further, in human beings, we have plausible biological explanations of the causal role by intention as *biasing*. I have argued that the biasing influence amounts to a type of cognitive penetration of relevant capacities in light of an agent's intention to act in a certain way (Wu, 2017).

Since the issue of automaticity and control will arise, let me repeat the basic idea here (for more, see my (Wu, 2013)). Control in action is rooted in the subject's intention to act in a certain way. To capture the nuances of skilled agency, I have proposed that we understand a subject's control *in action* as tied to the features of action that the agent intends. Put succinctly

> A feature F of an agent's action A is controlled iff A has F as a result of the agent's intention to A(F).
>
> A feature F of A is automatic iff it is not controlled.

The key feature of this analysis of automaticity and control is that it allows us to attribute both properties to any action. Every intentional

action will have automatic and controlled features. With these characterizations in view, it follows that for finite minds, actions will have mostly automatic features since the controlled features rely on the representational content of the biasing intention, and the intention is content-limited in that it cannot represent *every* feature of action. Every feature not represented by the intention counts as automatic. We can then characterize the acquisition of action capacities in terms of a shift in automaticity versus control, both over time and at a time, within and between subjects.

## 3.2 Attention

Returning to James' idea, we can see attention as emerging from solving the Many-Many Problem. Action is a specific input guiding a response, given what the agent intends. But the selective guidance by an input psychological state amounts to the agent's mind taking possession of one among many targets so as to deal effectively with it, namely to respond to that taking. But this is just James' description of attention, and the basic structure of selection for dealing effectively with what is targeted is echoed in so-called *selection for action* accounts of attention proposed by Allport (1987) and Neumann (1987) (see also (Wu, 2011a)).

Once attention has been located in a natural way, reflecting what James said that we all "know", its role is elaborated in the framework for action that draws on a behavior space. If one were to try to *point* to attention in the structure of a behavior space, when a path is actualized in action, one begins by pointing at the input that guides the response. Pointing to the guiding input state is just to indicate the mind's taking possession of something to deal effectively with it. In this picture, the agent's attending is rooted in the guiding of behavior (response). This Jamesian link allows us to use the structure of action to frame different aspects of attention as a psychological phenomenon.

I have been surprised that this account of attention has been resisted by many philosophers working on the topic. It seems to me that if one takes any of the science of attention seriously, where attention concerns the subject-level phenomenon as opposed to neural mechanisms, one must operate with the Jamesian conception. The reason is that all the experiments that philosophers of attention are likely to cite as investigating attention assume a basic methodological principle: one can study attention to some target X only if one ensures that experimental subjects are in fact attending to X during the experiment. Crucially, cognitive scientists ensure this by designing tasks such that to perform the task correctly, the subject must select the relevant X to inform their performance. This assumption implicitly reflects James' definition of attention, mentally taking possession of something to deal effectively with it. Scientific methodology reflects this conception.

In light of recent criticisms, I now recognize that the rubric, *selection for action,* is a misleading slogan. Some critics of the account have focused on the term "selection" to argue against attention as selection for action, specifically by emphasizing the temporal aspect of certain verbs (Vendler, 1957). Thus, one argument holds that attending is an activity while selection is an accomplishment (see Levy, 2019; Watzl, 2017). If one maintains that attending is selection for action, then the identification entails that attending and selecting should have the same relation to time. Yet activities are not accomplishments, for accomplishments come to an end in that they have a natural terminus, say running a four-minute mile, while activities like running so described do not have an end. Thus, attending need not have a natural terminus while selecting X does, namely when X is selected. Moreover, the processes described by activity verbs are homogenous in the sense that at any stretch of time in my running, it is true that I have run. This is not the case for accomplishments, for it is not the case that at any stretch of time in my running a four-minute mile, I have run a four-minute mile. Accordingly, while at any point where I am attending to X, I have attended to X, this is not the case for selecting X. This is true only at the end when I have reached my goal.

The argument is correct but it assumes that the slogan is the best way to capture the core of the theory. It is not. I cannot take those who have engaged with the view to task, however, given my ubiquitous employment of the term "selection for action" which invites a reading of the theory that construes it as conceiving of attention as more static and temporally restricted than I, and I suspect Neumann and Allport, actually conceive of it. There is a difference in emphasis here between myself and critics of the theory that arises from a difference in perspective. The argument against selection for action relies on conceiving of the selection for action account's slogan to be a bit of conceptual analysis. I have certainly offered definitions of attention, but the definiens is rooted in a *process* in which the agent partakes that is captured by the behavior space and by solving the Many-Many Problem, a dynamic process that takes time. Attention, and the idea of selection, are to be cashed out first in those terms. What I have learned from the criticisms is that the term "selection" or even "coupling", which I have appropriated from Alan Allport, do not grammatically fit the ideas fleshed out in understanding the process of acting. If you don't find the term selection apt, as the critics have emphasized, then drop the term, not the theory.

Let me acknowledge doing a sloppy job in respect of formulation and at times not respecting distinctions I have made. So, here's an attempt at being clearer, drawing on some passages from my earlier discussions. A better notion for understanding attention in terms of the specific input state that comes to be linked to a response when the agent acts is the notion of *guidance*.

> To speak of attention as selection for action is to speak of a way that the subject is attuned during action to relevant information such that it is deployed to inform the subject's response… action is constituted by a response guided by the agent's attunement to certain features of the world, including features of the subject him- or herself. There are, then, two necessary "aspects" of attention so conceived: (1) the attunement ("selection"), and (2) the link between the response and that to which the subject is attuned ("for action"). ((Wu, 2011a, p. 98), but see also the discussion in (Wu, 2014) around Figure 3.2, and the distinction between attention as state and as process in (Wu, n.d.))[3]

In respect of the non-deliberative Many-Many Problem

> [a solution] arises at the beginning of action and extends throughout its course, prompted by the demands of producing and guiding movement… skilled action requires that the subject be visually attuned to fine-grained spatial information and respond to it continuously and precisely over time
>
> (Wu, 2008, p. 1007)[4]

I hope that it is clear that the "link" in the first passage is what we have spoken of as a path in behavior space but that it involves a process of guidance of response by the input state as emphasized in the second. Again, I have not thoroughly analyzed the grammar of "guidance" since that is not my project. Attention's central role as guiding or informing the response is to be understood by conceiving of what is going on, psychologically and biologically speaking, in the agent. We must think through the lived life of action and not, merely, rely on conceptual analysis.[5]

The apparatus of the behavior space is to provide a framework for understanding human agency as a biological phenomenon of philosophical significance. The conception of action that invokes a behavior space is arrived at on conceptual grounds, but the theory of attention is the result of merging that account of agency with empirical and folk conceptions of attention. The surprising result was that this merger is fairly natural, beginning with James' definition and the echo of "taking possession… to deal effectively with things" in empirical methodology. Attention is then located within agency in a natural way. Consider visually-guided response. An agent sees many things but responds only to some of what she sees, say only to the visible X. It is then her seeing X and not her seeing other things, say some Y where $Y \neq X$, that guides her response. Her seeing X, in playing this role of informing her behavior, constitutes her mind's taking possession of X to deal effectively with it. That is, it constitutes her attending to X.

There is a thesis in the metaphysics of attention that I will simply assert here but which is connected to the issue of whether "attention

is cause or effect." Using the behavior space to depict the production of action, the issue is whether we should invoke attention as the process whereby a specific input comes to inform a response (attention as a cause, cf. the standard spotlight metaphor concerning visual attention) or whether attention is effectively explained by the input informing a response (attention as effect). I opt for the latter. There is no attentional spotlight in the sense just bruited. In another sense, attention does play a role in causation, namely, it guides response. If anything plays the role of the traditional spotlight, it is *intention*. That is, if the spotlight is a metaphor for what explains why a specific input rather than others informs response, then that is explained by the intention and not attention. Attention is set by intention in intentional action, and is essentially connected to the input states that are selected for action and which, in action, guide response.

What I have been calling the selection for action theory of attention strikes me as the best account of the phenomenon. The theory captures a core insight in James' description of our everyday grasp of attention as something we do, is incorporated into the methodology of experiments on attention in psychology and neuroscience, is accordingly invoked to interpret neural data as attentional (see Wu, 2014), provides a plausible computational theory of attention in Marr's sense of "computational theory", can do interesting philosophical and empirical work, and thus provides the frame for an integrated explanation of what attention is. *No other theory of attention has this broad set of constructive consequences and connections within philosophy and psychology.*

## 3.3 Action capacities

The idea of an *action capacity* is central to understanding movements of the mind. A capacity for action is one which can be repeatably actualized and when it is actualized, we have the agent's action. A behavior space is constituted by a set of actualizable action capacities that are available at a given moment in time given the relevant context. Such capacities are complex, constituted by integration of more basic capacities tied to the specified inputs and outputs. Thus, action capacities are partly constituted by perceptual, conceptual, and/or motor capacities and so on. It is the combination of these in respect of potential links in behavior space that constitute action capacities. Thus, when an action capacity is actualized, this involves the actualization of input capacities, say a perceptual capacity, that guide the actualization of response capacities such as conceptual or motor capacities.

In this way, an agent's behavior space is made up of what we can also call *abilities,* these being the capacities in question. Many of the abilities we have depend on what skills we acquired given a lifetime of acting, through learning, repetition, and practice. Learning involves the

acquisition of input and output capacities but also their being linked to constitute an action capacity. There are capacities that some agents have while others do not, leading to different behavior spaces. Thus, while a running back seeing a defender responds by juking, this coupling of a fine motor response to what the athlete sees is not reflective of an action capacity held by non-athletes even if on the basis of a similar perceptual state, non-athletes could move responsively in a different way to avoid a tackler, though likely not fast enough. Of course, young athletes try to emulate their favorite players and in doing so, try to acquire the relevant capacities. With appropriate learning and practice, they not only come to be more attuned to relevant features and sharpen motor responses, but learn to couple the two in order to perform the expert visually-guided movement.

The contrast between the skilled and unskilled agent points to the fact that we can use behavior spaces to explain the transition between unskilled to skilled behavior. What the skilled agent possesses is a learned expertise embodied in a specific action capacity acquired through training. An unskilled agent does not have that capacity, though she can have related capacities for movement. The purpose of training is then to shift the shape of a behavior space from links that are not the expert action capacity but which are developmental stepping stones to it, to a phase where that action capacity is in play but not yet quite expert-like, and ultimately to where that action capacity can be deployed in an expert-like manner, indeed automatically. Crucial to characterizing that transition is the shift in levels of automaticity versus control, something that we can track by looking at the subject's different intentions during the process of learning, training, and practice.

This diachronic perspective helps us to see how to individuate action capacities: they are capacities that *can be* targeted by an intention via *biasing* a type of "top-down" influence (see Wu, 2017). Put another way, action capacities are such that they can be actualized in intentional action. They constitute paths in a behavior space, and their actualization is necessary for certain intentions to have their contents satisfied. We can label this theoretical approach to explaining action capacities an *intentional-action-first* approach, one which identifies action capacities via their being potential targets of an intention. Accordingly, for human beings, we can highlight paths in a behavior space by identifying potential intentional actions, namely an action that can exemplify the agent's control in the technical sense defined in the previous section.[6]

Of course, all actions reflect a high amount of automaticity. There is conceptual space for *passive* actions in the sense of actions that are actualized action capacities independently of any intention. On my account, actions are passive when they are *fully* automatic, yet they remain actions because they are actualizations of action capacities, capacities defined because they are potential targets of an intention. Some will choose not

to call a behavior an action if it is passive in my technical sense, but I believe this imposes an unnecessary conceptual limitation of a theory of agency.

In the bodily action domain, utilization behaviors likely exemplify the passive exercise of an action capacity where bodily actions are simply triggered by the subject's environment, say when a mug is placed before a patient exhibiting utilization behaviors and they, automatically and without intending to, reach out to grasp it. If we conceive of this as just an odd reflex the patient has required, we could not speak of a direct defect to action capacities. It seems appropriate to speak of an alteration rather than a limitation of the agent's capacities for action. The issue for such patients is that their *agency* is defective and not that they have suddenly acquired a new reflex in the sense that doesn't exemplify agency but simply a mere behavior. Rather, the utilization behavior reflects a transformation of their agency, specifically an inability to inhibit inappropriate actions. The patient's abilities are wrung from her in a non-intentional, fully automatic action.[7] Similarly, in the mental action domain, obsessive thoughts in obsessive compulsion disorder, thought insertion in schizophrenia, or rumination on negative thoughts in depression exemplify behaviors to which subjects are passive. Categorizing these behaviors as actions, we recognize that these individuals retain their hold on the space of agency given in a behavior space. We understand them as agents, and talk of passivity of *action* underscores a specific and debilitating defect in agency via the automatic actualization of capacities for action. There is genuine suffering in automatized action.

There remains an issue of dividing mental from bodily action. Drawing the line sharply is difficult, but for current purposes, a coarse division suffices. An action capacity counts as *mental* when the output is a non-bodily capacity. For our purposes, we can treat bodily action capacities as those which when exercised involve a type of control of relevant muscles (this is to allow cases like going slack and staying there, standing at attention, or contracting a muscle as when one flexes it). Mental action will be understood as non-bodily action. There will be cases at the border. For example, motor imagery can involve motor preparation at the level of neural representations, but it will count as a mental action if this imagery does not influence the muscles. In fact, there may be physiologically borderline cases where such imagery does lead to small levels of muscle activity that precisely correlate with the content of the image. Whether it is important to make decisions about such cases is an issue I shall set aside. In what follows, I shall consider "pure" cases where such minimal influence is blocked.

The idea of a movement of the mind as a description of mental action is that actions involve traversals along paths in the behavior space, from input to output. In mental action, this traversal is best depicted in terms of intentional content. I shall discuss two movements in the final two

sections where content across time is transformed. In Section 3.4, I discuss a case of what we can call the shortest movement, namely covert perceptual attention, where the input state is modified in the output state. In Section 3.5, I discuss cases where the input state can be different from the output state. How one precisely draws a distinction between states that change and states that induce distinct states is a delicate matter that I leave open. For our purposes, the important thing concerns changes in intentional content as a way to track the progress of a mental action.

## 3.4 (Perceptually) attending as mental action

To understand attending as a mental action, I focus on the case of *covert perceptual attention*. The cocktail party effect provides the basic case. Covert attention is attention that does not require movement of the body, and in particular, movement of the relevant sense organ. So visual covert attention can shift even if the eyes are fixating on a point. Accordingly, covert attention is a mental action according to our rough and ready characterization. The contrast to visual covert attention then is visual overt attention which does entail that the eye has moved or at least is controlled to maintain contact on the target, and so is a bodily action. In humans, overt attention involves controlling the eye so that one is "foveating" the target of overt attention, the fovea being the area of the retina that allows for the highest spatial acuity.[8] Covert attention can occur outside of the area processed by the fovea.

The cocktail party effect is an example of auditory covert attention that is familiar from ordinary experience and inspired the first major experimental paradigm in the modern psychology of attention, the *dichotic listening* task. In dichotic listening, subjects are presented with two auditory verbal streams and are tasked to attend one of the two via instructions, the successful execution of which requires selection of one stream to guide parroting the heard words (again, recall the Jamesian methodological constraint in psychology). The dichotic listening paradigm is inspired by a familiar event: consider being at a cocktail party where you are in a conversation with your friend, your auditory attention focused on her words. Suddenly, from behind you, you hear someone say your name. This sound captures your attention, and you cannot help but listen in on that conversation. What are they saying about you? In paying auditory attention to the other conversation and thereby listening to it, you do not need to move your body. At the same time, you are losing auditory track of what your friend is saying, something that can lead to social embarrassment: "Hey, are you listening?" she asks suddenly, pulling your attention back to her. As this sort of experience shows, and was experimentally confirmed in experiments in the 1950s by Cherry (1953) and others, when you attend to one audible stream of words, what you are able to pick up in another stream is greatly diminished.

The cocktail party effect is familiar in ordinary life and exemplifies a way that we can exercise our agency without moving our bodies. We can furtively shift auditory attention at will between two conversations, and we can know that we are doing so, say listening to one's friend or to the group behind one. It is also phenomenologically vivid in the sense that if one pays attention to one conversation, one hears what is said at the expense of hearing what is said in the other conversation. In shifting attention, one flips the accessibility of the conversations.

Since action involves a traversing of the path between an input and an output capacity such that the input guides the output, covert attending as an action must exemplify the same structure. Consider then two inputs, each identifying the exercise of an auditory capacity of hearing a specific verbal stream. Without a coupling to an output, all we have in these inputs are auditory states that represent each voice. We can, however, attend to one rather than the other, and this leads to the phenomenally salient contrast in our accessibility to one voice versus the other as illustrated by the cocktail party effect. Given that such shifting of auditory attention is an action, what is its structure?

Focusing on the behavior space, we identify two input auditory states that are individuated by their intentional objects, voice 1, Sarah, and voice 2, Mark. In the behavior space that represents action possibilities, we identify these input states by abstracting them from our ongoing sensory experience of the world as distinct causal nodes (inputs), though, of course, these states are ontologically dependent on the overall experience. When one focuses on Sarah's voice rather than Mark's, then one's auditorily taking Sarah's speech must inform a response. Given the cocktail party effect, we can understand the response as an increase in one's attunement to Sarah's voice, say, to what she says. That is, the response is itself an auditory state whose content differs in its directedness to Sarah's voice relative to the initial auditory orientation to her voice. This difference in experiential uptake of Sarah's voice at the output stage is phenomenally salient, a type of focusing that is familiar from experience when flitting from one conversation to another, shifts in attention that we intentionally bring about. If we intend to listen to Sarah, we can do so with the attendant changes in our auditory experience. Similarly, if we intend to listen to Mark we initiate a different set of changes in auditory experience, an increase in how we listen to Mark at the expense of hearing Sarah.

When we attend to Sarah's voice, that auditory input state that counts as our taking in her voice guides a response which is just a change in that very state. Phenomenally, we have picked out Sarah's voice against other voices, and in intentionally attending to it, our phenomenal awareness is more focused or tuned in. The output, the sharpened auditory experience that we have, is a change in the nature of original auditory state, say in the fineness of grain of auditory access to (e.g., auditory representations

of) Sarah's voice. Thus, the input state provides the basis for changes in itself. We move from an auditory experience of lower auditory determinacy to one of greater auditory determinacy, something that we can lose the moment we shift attention, say when Mark utters our name and our attention shifts to his voice.

Attending covertly in this case instantiates the structure of action we have discerned: our initial auditory experience of Sarah's voice (input) alters in time to provide better access to her voice (output) because we intend to listen (attend) to Sarah. How we take Sarah's voice to be provides the basis of how we come to hear Sarah more clearly, in attending to her. In the act of covertly listening, the input and output are tightly linked in that the latter is a transformation of the former, thus constituting the act of attending. To the extent that we continue to listen to Sarah's voice in attending to the conversation with her, we sustain the attunement, something we can let go of if our attention is captured by another auditory stimulus (distraction).

Biologically, we have plausible models of how such intentional attunement works in covert attention, say in the sharpening of neural signals in respect of perceptible features. From the perspective of the theory of agency, however, what is involved is the traversal of a specific part of a behavior space and, accordingly, the actualization of a distinctive perceptual capacity that is subject to the agent's intention. Because this perceptual capacity involves the structure just noted, roughly that in listening intentionally, one's auditory state informs the development of its own greater access to the world, it is distinct from a *mere* perceptual capacity. There are, after all, ways to improve access to Sarah's voice that are independent of the posited action capacity. For example, one can reduce noise by lowering interfering volumes as when someone asks Mark to stop talking or when one turns down the music. The capacity for perception is not the capacity for perceptual attention though the capacity for perceptually attending does overlap in its basis with the capacity for perceiving. This is guaranteed by the conception of attending in play, namely one where the perceptual (input) state constitutes perceptual attention when it informs a response.

The issue of individuating capacities will need to draw on understanding the underlying biology. This is appropriate, for the argument in this essay only establishes the functional structure, as it were, of action via considering metaphysically necessary constraints on action. A priori reflection cannot on its own reveal the details of the biological implementation of action capacities. To shift modalities for heuristic purposes, consider the *premotor theory of visual spatial attention* which, in its strong, constitutive version, holds that visual spatial attention to L is just the preparation of a saccadic movement to L. Thus, when one covertly visually attends to L, there are typically measurable improvements in visual acuity at L (for a review of related work, see (Carrasco, 2011)).

As noted earlier for audition, improvements in acuity can be effected by manipulations of the outside world. Just as we can improve auditory acuity to a voice by reducing external noise or increasing the signal-to-noise ratio, we can improve visual acuity to features at a location by doing the same. But if the premotor theory is correct, the visual attentional capacity and the visual perceptual capacity, in respect of improved acuity, are distinct in part by having distinct realizations. For while improved visual acuity can be induced by changes in a stimulus, that achieved by the attentional capacity is an agentive capacity. On the premotor theory, this capacity is built on an underlying motor action capacity to move the eyes, one that can be intentionally deployed. Thus, tapping into the eye movement system directly can lead to improvements in visual processing (see the work of Tirin Moore and colleagues, especially (Moore & Armstrong, 2003)). As it turns out, the premotor theory is controversial and, in its constitutive version, possibly false (for a thorough dissection, see (Smith & Schenk, 2012)). I invoke it to remind us that there are deep biological issues that must ultimately inform our talk of action capacities and the individuation of kinds of capacities such as perceptual and attentional (action) capacities.

As I noted earlier, once we have the action capacity in view, identified by reflection on cases of intentionally acting, this allows for the possibility of fully automatic deployment of that capacity. When the capacity is passively actualized, we still have an action, just one that is independent of an intention to act. In the case of attending, the passive form is typically called *attentional capture*. So, hearing your name in the cocktail party will lead to the capture of auditory attention. In typical cases, attentional capture is overt, namely involving movement of the body, say the eye or ear towards a better orientation to the capturing stimuli. If there is covert attentional capture, then this will be the actualization of the action capacity that can be intentionally deployed in intentional covert attention. If the premotor theory of visual-spatial attention were true, then attentional capture would be the passive actualization of a capacity to program a movement of the eye but without actually producing movement. It is not clear that attentional capture is ever covert, but that is an empirical question.

We do not need to get embroiled in discussions of the underlying biology beyond noting their relevance in providing a full understanding of the nature of attending as an action. To emphasize the biology is to emphasize a check on limiting oneself to armchair theorizing about the nature of attention and agency. The picture in play is that covertly perceptually attending involves the deployment of an action capacity that identifies a very small movement of the mind in that the input state informs its output precisely because the output is the transformation of that very state.

Once we have in view covert attention as the result of an intention to attend so that one's better focus on a target is the result of an alteration

of an input state directed at that target, we can speak of the limiting case of maintaining attention on a target. In this case, the changes that are brought about by shifting covert attention are maintained, intentionally. That is, one keeps on a path in the behavior space because one intends to. When one gets distracted, attending shifts to a different target. This is a limiting case in that the content used to characterize the action can be the same throughout the time of the action. In that sense, this movement of mind is the smallest in that the intentional content does not change though we have an action in view only because there was an initial shift in content brought about by the intention to covertly attend. Maintaining covert attention is, perhaps, the mental action analog of standing at attention in bodily action.[9]

A final terminological point must be made regarding the term "attention." In this section, we are considering attending as a mental action, and as an action, it will have the complex structure that is revealed by reflection on the behavior space and the non-deliberative Many-Many Problem. When we speak of attending as an action, we are considering the actualization of a capacity that corresponds to a path in the behavior space. But we often isolate attention in the context of action, as when we speak of doing things in a way that depends on attention. Here, speaking technically, we need only have in view the input that will inform action. So, if we want to talk about attention as a component of action, such claims must focus on the role of the input and its function in guiding action. This is to focus on just part of the actualized path. Here, "attention" is used in a stative and not agentive sense, and underscores the distinctiveness of that state relative to other input states.

## 3.5 Mental action as attending

The idea that all mental actions are ways of attending follows naturally from the theory presented here, especially given its Jamesian heritage. Picking up on an earlier point, we can think of the state of attention as constituted by the input state where that state plays the functional role of guidance, the process of attending as constituted by the guiding of response by that input state, and the action of attending as constituted by the input state guiding response. So, if the auditory state that is intentionally directed to Sarah's voice is what guides response in covert auditory attention, then that auditory state constitutes the state of auditory attention. In that sense, the state of attention will have the intentionality of the constituting auditory state. For that reason, one can speak of attending *to Sarah's voice*.

In the movement of mind that is the sharpening of attunement to Sarah's voice, we can speak of attending to Sarah's voice, and this tracks the dynamics of the developing and sustaining of one's intentional directedness to Sarah's voice. In the context of solving the Many-Many Problem,

we have the traversal of a path in behavior space where this involves a type of selective intentionality. Adapting James' characterization of taking a "train of thought", the movement of mind involves a train of intentional directedness, here anchored on Sarah's voice. But this intentional directedness is in contrast to other paths in a mental behavior space that are constituted by a different train of intentional directedness. We grasp one train of thoughts rather than other potential trains of thought.

Some movements of mind are more expansive in terms of how they constitute different trains of intentional directedness. For example, a train of thought in James' sense might be a movement across different cognitive states individuated by their propositional content as one observes in propositional reasoning. Thus, in deliberating about a theoretical matter, one begins by thinking about a propositional starting point and then making an inference, arriving at a new thought individuated by its propositional content, and so on. That train of thought is one possible train of thought in the behavior space, and because of this, we have a selective intentional directedness across time as the agent deliberates. This is a form of attentional selectivity over time, a way of attending in thought. Similar points can be made about a sequence of visual imagery, recollection of a complex episode in memory, or mind wandering.

Thus, to the extent that mental actions involve transitions in states that are intentional where input states inform the output states, we have a type of attentional selectivity to intentional content, the sort of attention that can be described as "taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought." In that basic sense, all mental actions, individuated by sustaining or changing intentional content over time, are ways of attending to content.

## Notes

1. Thanks to Yair Levy for comments and discussion on the issues discussed in this paper.
2. I'll persist in speaking of a Many-Many Problem, but I now wish I had chosen the less gripping label "Problem of Selection" (Wu, 2013, pp. 250–251). As I've noted elsewhere, the Many-Many case is common and salient, but it does not cover all cases, one-many and many-one cases being obvious alternatives. There are also in effect one-one cases where the alternative is mapping an input to a null response (see discussion of pure reflexes). The Problem of Selection, which applies to all cases, is that action requires coupling an input to an output and for there to be action, this must come about. The solution to the Problem is a specific coupling among others, namely an action. The Problem of Selection, in the non-deliberative form, is answered by the agent's acting.
3. This passage also underscores what I now recognize to be a persistent problem with my previous discussions. Here, I use the "for action" to identify the linking of input to output that constitutes guidance. But "for

action" as a phrase doesn't imply any activity and hence, a natural reading of "selection for action" includes only (1). Phrasing aside, a link that is selected "is constituted by an input-output connection where the former guides the latter" (2011a, p. 100).

4. Having made clear two types of the Many-Many Problem, later discussions too often do not disambiguate these clearly. Here's an example that requires disambiguation (should readers venture back into previous publications of mine):

> The Many-Many Problem is illustrated by noting that to do anything at all at a time, selection of one among the four behavioral possibilities must take place within the behavioral space at that time. If selection does not happen, then nothing does. Thus, if there is to be action at this time, the Many-Many Problem must be solved: appropriate selections must be made whereby an input informs a specific output. Thus, a path in behavioral space is selected.
>
> (Wu, 2011a, p. 99)

The first solution is best read as concerning the deliberative problem but the second, which invokes selections whereby an input informs an output, concerns the non-deliberative problem.

5. This provides an answer to Levy's first argument against my position (this volume), the fact that selection seems to be over before the action is. I acknowledge that the term "selection" misleads, but the theory, when understood in light of the detailed structure and dynamics of the implementation of an action against a behavior space, identifies attention's role in the ongoing process of guiding response.
6. I set aside interesting questions about the development of the capacity for intentional action, as in how infants acquire such capacities. Much weight will be put on the fact that the relevant capacities that infants develop, say bodily movement capacities that become fine-tuned through the automatic movements generated by them, are potential targets for intention since the human brain has developed, in evolutionary time, to be an engine for intentional action. I would allow that some action capacities can be innate in that one is born with them, and these can be readily integrated with intentions when the subject is capable of issuing intentional action.
7. Yair Levy (2023, this volume) takes utilization behavior to be intentional in that they involve intentions-in-action and contrasts it with cases like doodling or drumming one's fingers. I'm not certain that there isn't an intention involved in utilization behavior, and there's a sense in which I think the point is biological and not conceptual. (I do think that intentions that yield action are intentions-in-action.) In any event, I can conditionalize the point. For all we know, this behavior can be bottom-up, driven just by the stimulus, the subject's seeing the mug. If it is, my point is that the behavior remains the passive actualization of an action capacity, and the reason for recognizing this is to recognize a defect in agency.
8. There is a complication here in that one theory of covert visual spatial attention takes attention to be tied to motor preparation to move the eye. This is the Premotor Theory of Attention. To covertly attend to a spatial location L involves preparing an eye movement to L, while to overtly attend to L involves moving the eye to L, to fixate on it. I have drawn the mental/bodily action divide in terms of whether or not the appropriate

muscles are involved, but no doubt there are issues here that this initial answer evades. I am not concerned with a precise demarcation.

9. Yair Levy (2023, this volume) raised a query regarding such cases which I take to be actions that are the "shortest" in distance in that the input and output states are in a way the same. Given the definition of control noted earlier, the relevant path is subject to the agent's control in that it is intended.

## References


Allport, A. (1987). Selection for action: Some behavioral and neurophysiological considerations of attention and action. In H. Heuer, & A. Sanders (Eds.), *Perspectives on perception and action* (pp. 395–419). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. doi: https://doi.org/10.1016/j.visres.2011.04.012.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, *25*(5), 975–979.

James, W. (1890). *The principles of psychology, Volume 1*. Boston, MA: Henry Holt and Co.

Levy, Y. (2019). Is attending a mental process? *Mind & Language*, *34*(3), 283–98. doi: https://doi.org/10.1111/mila.12211.

Levy, Y. (2023, this volume). The most general mental act. In Brent, M., & Titus, L.M. (Eds.). *Mental action and the conscious mind* (pp. 79–99). Routledge.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370–73. https://doi.org/10.1038/nature01341.

Neumann, O. (1987). Beyond capcity: A functional view of attention. In H. Heuer, & A. Sanders (Eds.), *Perspectives on perception and action* (pp. 361–394). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Smith, D. T., & Schenk, T. (2012). The premotor theory of attention: Time to move on? *Neuropsychologia*, *50*(6), 1104–1114. doi: https://doi.org/10.1016/j.neuropsychologia.2012.01.025.

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*(2), 143–160. doi: https://doi.org/10.2307/2182371.

Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford: Oxford University Press.

Wu, W. (2008). Visual attention, conceptual content, and doing it right. *Mind, 117*(468), 1003–1033. doi: https://doi.org/10.1093/mind/fzn082.

Wu, W. (2011a). Attention as selection for action. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 97–116). New York: Oxford University Press.

Wu, W. (2011b). Confronting many-many problems: Attention and agentive control. *Noûs*, *45*(1), 50–76. doi: https://doi.org/10.1111/j.1468-0068.2010.00804.x.

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the will* (pp. 244–261). Oxford: Oxford University Press.

Wu, W. (2014). *Attention*. Abingdon, UK: Routledge.

Wu, W. (2017). Shaking up the mind's ground floor: The cognitive penetration of visual attention. *Journal of Philosophy*, *114*(1), 5–32.

# 4 The Most General Mental Act

*Yair Levy*

## 4.1 Introduction

The time-honored philosophical project of *definition* – whether in the conceptual or the material mode – has by now seen its day. This is in no small part a consequence of perennial failures to define several key notions of human life and agency, such as *intentionality*, *knowledge*, *meaning*, *perception,* … More fundamentally, what has arguably driven the definitional project to its inevitable decline is the tendency in many quarters to conceive it in reductive or decompositional terms – as the project of identifying several more fundamental constituents which, properly conjoined, yield knowledge, intentional action, meaning, or whatever. Attempts to combine the various *analysands* in a way that would yield sufficient conditions for the *analysandum* have repeatedly come up against persistent stumbling blocks (exemplars include the deviant causal chains problem for causal accounts of action, reference, perception, etc., and the Gettier problem for analyses of knowledge). But the project can in fact be decoupled from the reductive aspirations with which many undertake it; it can look beyond reductivism to illuminate the phenomenon or concept of interest, and in this way arguably increase its chances of bearing fruit.

It is largely in this non-reductive spirit that the steadily expanding debate over how to define *attention* has been conducted. Its inclusion on the list of phenomena whose nature calls out for explanation would seem fitting, seeing as it displays the two features touched upon above: attempts to define attention have so far been inconclusive; and at the same time, attention's prominent place within the mental economy is very widely recognized. The present paper proposes to build on the latter feature in order to make progress with the former: it is argued that the proper way to capture the nature of attending is to conceive it as *the most general mental act there is* – to a rough first approximation, as the genus of which all other mental act-types are species.

This striking account of attention, set out and defended in Section 4.3, is novel in going against the grain of virtually all prominent extant

DOI: 10.4324/9780429022579-5

accounts, which work by purporting to identify the unique functional role of attention. The account is inspired by Timothy Williamson's account of knowledge as the most general factive mental state (Williamson, 2000, ch. 1). Some points of analogy and disanalogy with the source of inspiration will be outlined. At the same time, the explanation of attention as the most general mental act is animated by two striking (and no doubt related) pre-theoretical features of attention, referred to here as attention's *ubiquity* and its *heterogeneity*. These two features are described in the next section (Section 4.2). Before their import for the paper's constructive proposal can emerge, however, it is instructive to see first how they have led philosophers astray when developing their own accounts of what attention consists in. Section 4.2.1 rehearses some more or less familiar problems with one prominent account of attention, which is similarly motivated at least partly by the heterogeneity of attention – namely, Christopher Mole's "adverbialist" account. Section 4.2.2 then looks at a very different extant definition of attention – Wayne Wu's "selection for action" account – and explains why that account mishandles the other feature adumbrated, viz. the ubiquity of attention. This sets the stage for the more promising account of attention as the most general mental act in Section 4.3, which adequately captures both features.

## 4.2 The ubiquity and heterogeneity of attention (and how *not* to explain them)

We need first to get clear on what the ubiquity and heterogeneity of attention consist in. The former feature designates how widely attention figures in the different manifestations of our agency, in particular our *mental* agency. Quick reflection provides anecdotal yet highly suggestive evidence of how pervasive attention actually is in our mental lives. Thus consider looking at some object, listening, reading, reciting a poem, deliberating, performing a calculation in one's head, judging, deciding …[1] Each of these act-types (and a host of others that could readily be added to extend the list) seem clearly to entail attention. It would be impossible to perform any of them without paying at least some minimal degree of attention to the object one is acting upon (deciding *to V*, reciting *R*, judging *that p*, etc.). The *heterogeneity* of attention consists in the various different forms attention can take – involving different modalities, implemented by different mechanisms, and performing different functional roles.[2] Thus *perceptual* attention is variously instantiated in all sense modalities (looking, listening, smelling, etc.) And attention in *thought* occurs whenever one performs some specific type of content-manipulation as part of one's thinking (problem-solving, conjecturing, mental arithmetic, recollecting, …). Attention is additionally, and somewhat differently, involved in *executive* mental functions, such as deliberating and deciding.

These very different manifestations of attention inspire very different conceptions of the *mechanism(s)* that implement the paying of attention, as studied by psychologists. Thus, for example, psychologists endorsing the "feature-integration" paradigm have claimed that attending is to be understood as the process of integrating different features of a perceived object (Treisman, 1993; Treisman & Gelade, 1980). Others portray attention as a 'filter' that selects stimuli for later processing in the face of capacity-limitations or informational "bottlenecks", which prevent the processing of all incoming stimuli (Broadbent, 1958; Lavie & Tsal, 1994). Still others see it as a process that biases the competition between rival incoming stimuli, favoring one over the other and thereby once again helping the system cope with limitations to processing capacity (Desimone & Duncan, 1995); or alternatively again, attention is construed as a moving 'spotlight' of sorts, which selectively allocates cognitive resources to available stimuli on the basis of their spatial location. And so on.

The above illustrates, in a quick and dirty way, the twin features of heterogeneity and ubiquity of attending. Any theory of the nature of attention worth its salt cannot fail to be responsive to both features. But an overblown response is equally damaging, as we shall now begin to see.

### *4.2.1 Overreacting to heterogeneity: Attentional adverbialism*[3]

The psychological research program investigating the nature of attention is widely regarded today as ill-suited to providing a fully general account of what attention consists of. Some empirical paradigms were briefly cited above; each faces decisive counterexamples if read as attempting to capture the nature of attention as such. Thus, for example, the feature-integration theory fails to account for cases where one attends to a single property of some object (say, its color). Similarly, the filter paradigm is ill-suited to capture scenarios where only one stimulus is available for processing and hence no competing stimuli need to be screened off. And again, the spotlight theory for its part fails to account for cases where attention is not allocated on the basis of location – e.g., attention in thought.

The suspicion that empirical accounts of attention fall short of accounting for the fully general phenomenon in all its different guises is nicely captured in an oft-quoted dictum expressed by Allan Alport some 30 ago:

> [E]ven a brief survey of the heterogeneity and functional separability of different components of … attentional control prompts the conclusion that, qua causal mechanism, *there can be no such thing*

> *as attention*. There is no *one* uniform computational function, or mental operation (in general, no *one* causal mechanism), to which all so-called attentional phenomena can be attributed.
>
> (Allport, 1993, p. 203)

Attentional adverbialism, as the view is dubbed here, provides a diagnosis for Allport's suspicion that the psychological research program into the nature of attention is in principle unable to identify a unique mechanism that corresponds to attending as such. According to adverbialists, extant psychological theories fail to identify the mechanism or process attention consists in because they operate under the misguided assumption that attention consists in some process or other, when in fact it does not consist in any process whatsoever. Rather, attending is an adverbial phenomenon: It consists in a way or manner or mode some process might occur – viz., the *attentive* manner or mode. On this view, virtually any process can instantiate attention if it is performed in the right circumstances or in the right way, that is, attentively. (To illustrate, compare *employment*: virtually anything that one does could count as one's employment if it is done in the right circumstances – i.e. if one is contracted to do it, is compensated for it, and so on. Hence, employment is a paradigmatic adverbial concept.)

This striking view of attention has historical precedents in Alan White (1964) and F. H. Bradley (1886).[4] But its most systematic development is due to Christopher Mole (2011). One of the first tasks Mole's adverbialism faces is of course that of specifying what exactly the attentive manner involves. If 'attending' does not designate any mental act or process taking place but rather a way in which some acts and processes take place, what does this way actually consist of? In slogan form, Mole's answer is: 'Cognitive unison.' Very roughly, the idea here is that that one is attending to some task *t* iff all of one's available cognitive resources are devoted to executing *t*. Now as some others have noted, and as Mole himself is fully aware, this proposal seems to land him in a problematic eliminativism towards partial or divided attention. For if *V*-ing attentively requires that *all* available cognitive resources be directed at the task of *V*-ing, it is unclear how we can make sense of the (intuitively obvious) gradeability of attention. (For discussion of this point, see Mole (2011, pp. 83–85), Wu (2014, pp. 102–103), and Koralus (2014, pp. 42–44).)

A second challenge for Mole's adverbialism concerns what may be called 'intrinsically attentional deeds.' The idea that attention consists in a manner or a mode of performing some act requires of course that there be also a *non*-attentive manner of performing the act in question. Otherwise, there would be no way to distinguish cases where attention is, from cases where it is not, being paid. Now, this constraint does seem to be satisfied for plenty of *bodily* act-types, e.g., driving, painting, sawing, and cooking: Each can be performed either attentively or inattentively.

However, at least some mental acts are *intrinsically attentive*; their performance necessarily implies that the agent is paying attention to at least a certain degree. To see this, recall some of the examples cited above when introducing the ubiquity of attention – looking at the car crash, smelling the lilies, doing some mental arithmetic, and so on. It is hard to make sense of the thought that one can perform such acts without paying *any* attention whatsoever (though no doubt one can perform them with more or less attention).[5] But if there is no non-attentive manner of performing these acts, then, of course, there can be no attentive manner either, and hence no manner for attention to consist in.[6]

A further objection to attentional adverbialism is the most significant for present purposes. This objection turns on the suspicion that the view is too quick to infer from Heterogeneity that attending is an adverbial phenomenon. As the counterexamples cited above illustrate, none of the subpersonal mechanisms identified by psychologists plausibly constitute attending as such. Each proposed mechanism is vulnerable to counterexamples which imply that it is not necessary for attending:

> There can be attention without feature-integration, attention without biased competition, attention without filtering, … More controversially, Mole suggests that (something like) Heterogeneity also implies that no proposed mechanism is *sufficient* for attending
>
> (Mole, 2011, pp. 36–41).

However, to infer unnecessity or insufficiency from Heterogeneity is to overlook the possibility that a more general process, pitched at a higher level of abstraction than any subpersonal cognitive mechanism, *could* successfully capture attention in all its different guises.[7] Perhaps, for example, attention consists in the pairing of some incoming stimuli with a behavioral response (a proposal examined in Section 4.2.2), or alternatively in the structuring of mental content, as Watzl (2017) argues. In other words, perhaps filtering, integrating, spotlighting and the like can all be subsumed as instances of a more general-level functional role that *can* successfully capture attending in all its heterogenous glory. Mole's argument fails to address this possibility.

### *4.2.2 Overreacting to ubiquity: Selection for action*

We have seen how Mole's attentional adverbialism overreacts to the heterogeneity of attention, and hastily concludes that attention is an adverbial phenomenon that consists in cognitive unison. A different theory of the nature of attention, the so-called 'selection for action' theory, is informed by the other feature of attention noted earlier, viz. its ubiquity. But similarly to adverbialism, this view also constitutes an overreaction to its sensible point of departure, which consequently prevents it from issuing in an adequate account of attention.

The 'selection for action' view attempts to explain attention by building upon attention's supposed functional role in facilitating action. One of the main advocates of the view is Wayne Wu (2008, 2011, 2014), who is inspired by the ideas of Allport (1987) and Neumann (1987). Central to Wu's development of his version of the 'selection for action' view is the thought that agents face what he dubs the 'Many-Many Problem' (echoing Allport's 'many-many possible mappings'). According to Wu, action is only possible against the background of a behavioral space comprised of various possible acts which could be performed in response to various different incoming stimuli. Each agent thus faces the problem of selecting which input to couple with which behavioral response:

> Here is one way of raising [the Many-Many Problem]: how is coherent action possible in the face of an overabundance of both sensory input and possible behavioral output? Much of the input is irrelevant to the agent's current goal, much of the output incompatible. Action arises only if the agent reduces this many-many set of options to a one-one map.
>
> (Wu, 2008, p. 1006)

To illustrate the Many-Many Problem (henceforth, MMP) and its one-one solution, consider a simple action of kicking a ball (Wu, 2014, pp. 79–80). To simplify, suppose the agent receives just two incoming sensory inputs: The sight of a basketball and of a football, and has available two possible responses, kicking with her left foot and kicking with her right foot. The agent solves the problem when she couples one input with one response or in other words, when she *acts* – kicks the football with her left foot, say. According to Wu, this selective matching of input to behavioral response is achieved by, and indeed amounts to, attending. In this way, attention facilitates all action – both bodily and *mental* (Wu, 2011, pp. 99–100).

The 'selection for action' view builds upon the ubiquity of attention by defining attention in terms of its purported necessity for agency in whatever guise, a necessity which in turn readily explains the pervasiveness of attention in our active lives (even though, importantly, the nexus between agency and attention posited by the view is considerably more expansive than Ubiquity above allows – encompassing not just mental, but also bodily agency; more on this in Section 4.3). Nevertheless, once again, the resultant account of attention misconstrues attention's connection to agency, and consequently falls short. For we can and often do attend without solving the MMP.

To see this, consider first one's act of reading some news item in the morning paper. To begin with, the case can seem congenial to the idea that one successfully negotiates an MMP here. For example, we may suppose that there are different items one glances in the paper spread out in front of one, and different means or ways one could choose to

read each item (e.g., to read it silently, out loud, etc., etc.) Thus when one chooses to silently read the item on p. 17 about a Chihuahua who saved the pet hamster from drowning in the toilet, plausibly enough an MMP has indeed been solved and attention has been paid.

However, the correspondence between solving the MMP and paying attention seems to last only for as long as it takes to complete the selection which solves the problem. And crucially, attention can and often does continue to be paid well after any such selection has been completed. For example, the agent in the case described may well continue reading the Chihuahua item even when her action no longer presents any MMP. For the MMP has been solved once a particular stimulus has been coupled with a particular response, i.e., once the agent has chosen which item to read and how exactly to read it; but her reading – and hence, her *attending* – goes on beyond that point, till she reaches the end of the news story. This is an untenable temporal mismatch between what are purportedly equivalent processes.[8]

To explain away the appearance of mismatch, Wu must insist that selection for action actually persists up until the agent finishes reading the news item. Two ways in which this might happen suggest themselves; neither sits well with the idea of an MMP as stated. The first way turns on the claim that selection for action continues to occur *successively* at each moment throughout the episode of reading and not just when one initially confronts the selection problem. But viewing one's action in these terms is contrived. Suppose, as will often be the case, that one's solution to the original MMP remains unchanged throughout one's action. That is, one does not change one's mind as to which item to read, how exactly to do so, etc. at any point while reading. It would be very awkward to then maintain that one must nevertheless constantly 'reaffirm' as it were one's initial choice to read *this* particular item in *that* particular way. The suggestion is both at odds with the phenomenology of such ordinary acts, and oddly wasteful of cognitive resources as well.

The second way Wu could attempt to defuse the temporal mismatch is to claim that selection for action is required to *sustain* the agent's attention beyond her initial solving of the MMP. However, this seems to put a rather different gloss on what 'selection for action' means in the present context. For sustaining an executed coupling – whatever exactly this turns out to involve – is certainly *not* a matter of coupling some stimulus with some response out of multiple alternative pairings. A quick way to see the difference is this. To couple a (previously uncoupled), stimulus/response pair is to cause a change to occur. To sustain an existing coupling, on the other hand, is to uphold the status-quo.[9,10]

Moving on from cases of temporal mismatch, we turn next to another range of scenarios that further undermines the necessity of solving the MMP for attention.[11] These are cases where a subject pays attention but has no range of stimuli or responses to select from. Call them cases of

'degenerate selection.' The claim is that degenerate selection is not tantamount to selection for action.

One illustrative type of degenerate selection scenarios involves quasi-sensory-deprivation. A blindfolded, gagged, ear-muffed, glove-clad subject can detect nothing but a faint odor of eucalyptus in his environment. There is only one stimulus he could respond to, and only one possible response available to him. Yet surely he can attend to the smell. Or consider a pain sensation so acute that it effectively screens off all other incoming stimuli. Once again, attending to the pain is clearly possible, even when no alternative course of action is available and so no MMP presents itself.

Wu is aware of the potential threat degenerate selection poses to his view (Wu, 2014, pp. 81–82, pp. 89–90). In response, he insists that there is still selection for action even in "a putative *one-one* behavior space so long as the action is something that need not be done, so in effect there remain two options: to act or not to act" (2014, p. 81). But this expansive conception of solving the MMP is strained. In ordinary cases of the MMP, the question facing the agent is 'Which act shall I perform?' – i.e., which stimulus to act on, and in what way. *That* question plausibly presupposes that the agent acts in some way. In cases of degenerate selection, in contrast, the question becomes: 'Shall I act?' And this is a very different question, one that is typically informed by rather different considerations and which does not presuppose that action takes place at all.

One way to bring out the divergence of the two questions is to recall the puzzle that animates the thought that attending requires solving the MMP, as quoted above: 'How is coherent action possible in the face of an overabundance of both sensory input and possible behavioral output?' (Wu, 2008, p. 1006). *This* is the puzzle 'selection for action' was recruited to address; the answer it provides is what lends the view its initial plausibility. Now to be sure, an *overabundance* of inputs or outputs is not essential for the puzzle to be sensibly raised. Indeed, even one-many and many-one spaces may still give rise to a version of the MMP. Crucially, however, this is *not* so with a degenerate one-one space. For when the latter obtains, the possibility of coherent action becomes entirely *un*puzzling: action cannot fail to be coherent in cases where there is just one stimulus and just one response available. Hence, if degenerate selection poses any relevant question or problem at all, it is one that can hardly be regarded as a version of the MMP.

In response, one may wish to deny that the MMP in fact presupposes that the agent acts in some way or other. Put differently, one may insist that "none" is an intelligible response to the question "which act shall I perform?", as posed by the MMP. But this does not help. For even if inaction *is* in some sense an acceptable solution to the problem one faces, *it is anyway not a solution that involves attending*, since one could fail to do anything whatsoever – including to attend – when implementing it.

Consequently, solving the MMP in degenerate scenarios does *not* guarantee that attention is being paid. The selection for action view is extensionally inadequate.

The ultimate aim of this critical section has been to clear ground for the constructive account of attention as the most general mental act, defended below. Two central features of attending which motivate the account – its ubiquity in mental agency and its functional heterogeneity – have been emphasized. And two prominent alternative accounts to the one endorsed here, which are similarly informed by at least one of these features, have been shown to come up short. The 'selection for action' view is unable to handle scenarios of degenerate selection,[12] as well as scenarios where attention is being paid well after the MMP has been solved. And attentional adverbialism for its part fails to leverage the point that none of the processes identified by psychologists seems equivalent to attention in order to establish its much more radical claim, whereby attention does not consist in any kind of process whatever. Moreover, adverbialism is incompatible with intrinsically attentional deeds. These failures are instructive inasmuch as they point towards more promising avenues forward, as we shall see in the next section. Attention can be adequately explained in a way that respects both Ubiquity and Heterogeneity.

## 4.3 Attention as the most general mental act

### *4.3.1 The entailment of attention by mental action*

If, contra attentional adverbialism, there *is* some process that takes place whenever attention is being paid, then a good place to look when trying to characterize the process(es) in question would be the class of intrinsically attentional deeds (a class which, the reader will recall, adverbialism fails to accommodate). For these are mental processes that invariably imply attending. They include watching, listening, smelling, concentrating, and other basic agential operations of our sensory and cognitive faculties. Here is Alan White on this point (incidentally making his view hard to reconcile with what seems at other places to be a clear endorsement on his part of adverbialism):

> Because we focus on what is perceptible by using the appropriate sense-faculty and on what is intelligible by making it the object of our thinking, we can specify the general notion of *attention* in terms of these particular perceptual and intellectual activities … when we speak of attention being paid or given, drawn or attracted, it is basically some set of these perceptual and intellectual activities to which we refer.
>
> (White, 1964, pp. 7–8)

But attention is not regarded here as restricted to the *basic* employment of sense organs or thought processes, as in smelling a flower or concentrating on a difficult idea. The claim is the more ambitious one, that every instance of mental action whatever implicates attention:

> [*Entailment*] For a mental act *V*, 'A *V*-s [*preposition*] *O*' entails '*A* attends to *O*.'

*Entailment* is put forward here (somewhat speculatively) as a datum any reasonable theory of attention should aim to explain. The explanation proposed here will emerge after a bit more groundwork is completed, explicating and defending *Entailment*.

The extended view of attending as involved not only in basic active perception and thought, but more broadly in mental agency as such, is supported by two principal considerations. The first has already been mentioned: The range of anecdotal yet highly suggestive corroborating evidence coming from reflection, including performing mental arithmetic, recollecting, reciting poetry, hypothesizing, visualizing a scene, reading, daydreaming, deliberating, … The list could readily be extended, apparently constrained only by one's view on the scope of mental agency. The second consideration is simply that all the sophisticated mental acts ultimately involve some form of the more basic (active) perceptual or intellectual modes, i.e., the intrinsically attentional deeds – listening, looking, etc. Thus, reading involves looking at sentences on the page, solving a puzzle involves focusing one's intellectual attention on the puzzle, and so on.

One might protest that reflection raises not only confirming but also refuting instances. Consider *deciding*, for example. Aren't some decisions at least momentary and "spontaneous", and hence executed without the deciding agent paying attention? The scope of mental agency is a persistent point of contention in contemporary debates, with decisions representing one controversial case (See, for example, Mele, this volume). Those who deny that deciding is acting at all will have no problem with the objection. Those who affirm the claim will anyway tend to view deciding as a more extended process which encompasses more than the instantaneous plumping for one course of action over another. For only such a temporally expanded conception can make deciding seem like it has the proper temporal profile to count as a genuine action, since actions are never instantaneous but rather unfold over time.[13] Now this temporally expanded conception must presumably include at least some of what takes place prior to the moment of decision in the more narrow, instantaneous sense, such as weighing the different alternative courses of action. And once that much is included, the place of attending in decisions starts to come into view, resembling its place in episodes of deliberation. A parallel treatment confirms that *calling to mind* is likewise not a

counterexample to the present association between attention and mental action. A suitably broad understanding of what one is up to when calling some fact to mind reveals the place of attention in such episodes.[14]

In some cases, however, it is precisely a broad understanding of one's action that may seem to threaten *Entailment*. Consider, for example, a case where one is solving some difficult mathematical puzzle. It may be that, at certain moments throughout the period in which one is solving the puzzle, one is also doing something else mentally – say, planning what to prepare for dinner tonight – with one's solving of the puzzle being suspended in the background, as it were. In those moments, the thought goes, one may still be correctly described as solving the puzzle even though one is not just then attending to the puzzle but rather to one's dinner plans. The fix for this apparent problem is simple: it is to note that '*V*' as it occurs in *Entailment* above should be restricted to the more narrow, localized sense which pertains only to the mental process(es) one is engaged in "directly" (or those taking place in the "foreground"). Henceforth, this qualification to *Entailment* will be taken as read.

*Entailment* identifies the object of action with the object of attention, *O*. The appropriate substitutions for *O* will depend on the type of attention at issue. In episodes of perceptual attention, '*O*' stands for the object or property perceived. With attention in thought, in contrast, '*O*' stands for some abstract entity – an idea, puzzle, question, consideration, and so on. In some cases of action in thought with *O* as its object, attention to *O* will be paid in a roundabout or indirect way. Consider for example the act of deliberating whether to go out to the local theatre tonight to see the new film on show. In so deliberating, one attends to the reasons for and against going – e.g., that the film has received mixed reviews, that one's friends will be there, and so on. Nevertheless, the primary object of one's attention is the question whether to go see the film; attending to the reasons that bear on the question is one's way of attending to the question itself.[15] Similar points apply to other mental actions on *O* which may likewise raise doubts as to whether their performance entails attending to *O* in particular. Conjecturing or hypothesizing that *p*, for example, involves among other things considering, and hence attending to, possible implications of *p*. And once again, doing so is a way of attending to *p* itself.

The implication of attending in exercises of mental agency is a prevalent and time-honored idea. Some, such as William James, go even further in supposing that attention is essential to action *as such*, mental and bodily – at least, voluntary or intentional action.[16] Malebranche (1997) argues that attention is essential to free action. And the 'selection for action' view developed independently by Allport, Neumann, and Wu also seeks to exploit this connection, as we have seen. However, the suggestion that attention is essentially connected to both mental *and bodily* action is not endorsed here. To see why, recall *Entailment* again.

And consider an attentive performance of some bodily act, for example driving a car. The attentive driver, who drives with attention, does such things as look out for pedestrians crossing the street, identify the location of a wailing siren, work out which exit leads to her destination, and so on. She is driving the car with attention. Now with some flexibility, perhaps it could be claimed that the object of her attention is ultimately or fundamentally *also* the object of her action, viz. the car (Is the car approaching the pedestrian crossing too fast? Is it heading in the right direction?) Be that as it may, however, the performance of the *in*attentive driver certainly does not vindicate the identification of the object of action with the object of attention. For *this* driver need do none of the things the attentive driver does. The *car*, therefore, may not be an object she is attending to. Evidently, the case is one where the agent performs a bodily act on some object (viz., the car) without attending to it.

An opponent might insist that the inattentive driver too must be paying attention to some object(s), even if not to the car she is driving. If this point is incompatible with *Entailment*, the thought goes, it is the latter, not the former which should be called into question. Perhaps both mental and bodily action entail attending, while in the case of bodily action at least, it need not be attending to the object of one's action specifically.

The claim that bodily action necessarily implies attention in some form is contentious (must one necessarily be attending to anything at all while wandering aimlessly through the woods, one's mind being in a passive meditative state?) But even assuming the claim should be accepted, far from refuting *Entailment*, it seems in fact ultimately to be explained by it. For if the inattentive driver is attending to something other than the car she is driving, it is precisely *because* she is performing some mental act(s) which does conform to *Entailment* – e.g., listening to the radio, working out what to pick up at the store on the way home, etc. Hence, the restriction of *Entailment* to mental action is not problematic.[17]

### *4.3.2 From entailment to determination*

Assuming the entailment of attention by mental action as stated in *Entailment* is accepted, how might one explain it? Broadly speaking, there are two main strategies available. First, attention could function as a precondition or prerequisite for mental action. The idea would be that in order to perform some mental act on *O*, one must first attend to *O*. Alternatively, *Entailment* may be explained by some sort of general/specific relation that obtains between attention and mental action. For example, attention might be a genus of which all mental acts are species. It is the latter option that is endorsed here, yielding the proposed account of the nature of attention:

> [*Most General*] Attention is the most general mental act. If one is attending to *O*, one is performing some more determinate mental act on *O*.

*Most General* is inspired by the Williamsonian account of knowing as the most general factive mental state (Williamson, 2000, ch. 1).[18] The account construes all factive mental states as species of sorts of the genus 'knowledge': Remembering, perceiving, realizing, detecting, and so on, are all, according to Williamson, more specific kinds of knowing. On the present account of attention, the claim is that all mental acts are more specific kinds or ways of attending. How this thesis explains *Entailment* should be immediately apparent. What takes a bit longer is clarifying what exactly it means.

Williamson's account may be similarly motivated by the explanation it provides for the parallel entailment of knowing by all other factive mental states. Here also, two similar possibilities suggest themselves for how the explanation might go: Either we should invoke Williamson's claim that knowing is the most general factive state; or we could try to show that knowing is a precondition for being in any factive state whatever. Now Hyman (2014) argues forcefully that the latter is the correct explanation, at least with respect to a wide range of factive states. Being glad that *p*, disliking the smell of *p*, regretting that *p*, being amazed that *p*, hating the fact that *p,* etc., etc. are all, according to Hyman, counterexamples to Williamson's claim. They are all factive states, but they cannot be understood as ways or modes of knowing. Rather, knowing precisely stands as a precondition or prerequisite for being in those states. In order to be glad that *p*, for example, one must first know that *p*; but one's being glad is not the way in which one knows.[19]

The aim here is not to settle the dispute between Williamson and Hyman. But it is instructive to notice that Hyman's alternative picture of how factive mental states relate to knowing is not a workable basis for an explanation of *Entailment*. Looking, listening, recollecting, deliberating, deciding, imagining, hypothesizing, reciting, calculating, and so on – none of them have attending as a precondition. It would be exceedingly odd to suggest that one must first attend to *O* in order to be in a position to look at it or call it to mind or whatever. In fact, looking, recollecting and the other mental acts seem precisely to *constitute* specific ways or modes of paying attention – ways individuated by the particular perceptual capacity or the particular mode of thought they deploy. Rival accounts of attention that reject *Most General* face the challenge of coming up with an alternative explanation for *Entailment*. With the precondition strategy ruled out, their task becomes daunting.

*Most General* is framed in way that is deliberately neutral between two different readings of the precise metaphysical relation that obtains between attention and mental action: attention could either be the genus

of which all mental acts are species; or alternatively, it may be a determinable of which all mental acts are determinates. The most conspicuous difference between the two alternatives is that only on the genus-species relation can the differentia or distinguishing feature of the specific be conjoined with the general to single out the specific. For example, *equal-sided* may be conjoined with *parallelogram* to single out the specific *rhombus*. But no such conjunction could ever specify a determinate in terms of its determinable. Red is famously a determinate of the determinable colored. And red is a way of being colored rather than something *in addition* to being colored; the only property that could be conjoined with colored to single out red is red itself.

Picking up the slack, two considerations favor opting for the determinable/determinate over the genus/species as the relation obtaining between attention and mental action.[20] First, as noted, the genus/species relation does, and the determinable/determinate relation does not, allow for an analysis of the specific in terms of the general property. And in fact, no such analysis of specific mental acts in terms of attention is available. To verify, consider what might be conjoined with attending to yield *listening*. A natural first response is "using one's ears." Perhaps one's listening to *O* could be analyzed as one's attending to *O* by using one's ears. However, the proposed analysis fails to provide a sufficient condition, since not any old use of one's ears while attending amounts to one's listening. For example, if one *looks* at *O* through one's glasses which are held in place by one's ears, there is a sense in which one attends and does so by using one's ears. Attempts to sharpen the apposite way of using one's ears all lead back to 'the listening way.' Nor can we say that listening is attending to *sound*. For one can attend to a sound without listening to it, e.g., by calling it to mind.

A second (and related) reason for thinking that mental acts are determinates rather than species of attending is this. The fact that species can be analyzed in terms of their genera reveals the metaphysical or explanatory *priority* of the latter over the former. On the other hand, the determinable/determinate relation exhibits the reverse order of priority: *O* is colored by being or because it is red, not red by being or because it is colored. And the present context displays the same priority structure: *A* is attending to *O* because she is listening to *O*, not the other way round. Hence, if attention is indeed the most general mental act, it plausibly stands to all other mental acts as a determinable to its determinates, not as a genus to its species.[21]

> [*Determination*] Attention is a determinable of which all mental acts are determinates.

As stated, *Most General* and *Determination* do not distinguish between different kinds of mental act; they claim that any mental act-type

whatever constitutes some determination of attending. This deliberately allows the present account to cover both intentional and non-intentional mental act-types. To illustrate the variety falling under the latter category consider absent-minded behavior; habitual actions performed on 'auto-pilot'; as well as such acts as sliding one's hand on the wall as one walks down the corridor, idly drumming one's fingers, fidgeting, and so on. Unlike intentional acts, such acts are often performed with no particular purpose or aim that one is trying to accomplish, and without any intention in mind. Nor are they typically done fully *consciously*: one may catch oneself pulling a face or shifting position.

The distinction can help explain away some apparent counterexamples to *Most General*. Not only bodily but also some mental acts belong in the non-intentional category – for example, talking silently to oneself, gazing idly into the distance, skim-reading billboard signs as one drives past them, and so on. Now such behavior seems to involve at least a minimal degree of attention (as an indication of this, notice that one is often able to report at least some of the information on the billboards one glances, or detect some objects in the scene one is gazing). It can therefore be thought to raise counterexamples to *Most General* in the form of attention without mental action – but only if one wrongheadedly equates action with intentional action, overlooking the non-intentional. Cases where one pays no attention at all, and displays no conscious awareness of one's mental goings-on, as when a solution to a problem that has been bothering one suddenly occurs to one, are likewise not a source for counterexample to *Most General* or *Determination*. For in such cases, one does not seem to be acting mentally, either. Rather, one is experiencing a subconscious mental process that results in the solution occurring to one.

Some other cases call for a somewhat more delicate treatment. It is widely held that our attention is sometimes captured rather than intentionally or voluntarily given – as for example in the famous 'cocktail party effect', where the mentioning of one's name grabs one's attention; the sound of a loud siren will tend to have a similar effect. As such, attentional capture (a form of what is sometimes referred to also as 'bottom-up' attention) may seem to demonstrate the possibility of attending as patients, not as agents, contra the present account. The threat such cases pose to the present account is neutralized once we carefully tease apart the actional from the non-actional, and the attentional from the non-attentional, in episodes of attentional capture. The mere *registering* of a cognitive or perceptual input such as a loud siren does not yet indicate the presence of attention. It is only after the input has been registered – when one starts to think about it, listen to it, or whatever – that attention is being paid. Thus, episodes of attentional capture can be usefully broken down into a pre-attentive stage and an attentive stage. And this division in turn corresponds to one between a passive and an active stage.[22]

As further evidence that attentional capture is indeed bifurcated in the way proposed, consider a standard measure for whether some piece of behavior is agentive or not – namely, whether the subject has *control* over what she is doing. If the pre-attentive stage of attentional capture is also non-agentive as suggested, we would expect the subject to lack the capacity to control it, and vice versa for the attentive stage. And in fact, this is precisely what we find. In the cocktail party effect, one cannot help but hear one's name uttered when it is first received as input. But thereupon, one can decide to stop or continue listening to the conversation.[23]

This section has made a preliminary case for understanding attention as the most general mental act. It has defended the claim that acting mentally entails attending, and has proposed to explain this phenomenon – and thereby, to explain attention itself – by seeing particular mental acts as determinations of attention. The result is a simple and rather elegant account of attention. But some readers may find it disappointingly deflationary or thin. How much insight does it actually provide into the nature of the phenomenon?

The reader should not lose sight of the fact that thinking of attention as the most general mental act represents a stark alternative to what is by far the predominant approach to theorizing about attention. Most extant accounts of the nature of attention work by proposing theories of attention's *functional role*. Mole's idea that attention achieves cognitive unison and Wu's suggestion that attention selectively matches inputs with responses are just two prominent examples.[24] The account defended here, in contrast, explains what attending is by illuminating its central place across the (active) mental economy, rather than by assigning to it any *specific* cognitive or behavioral role. Furthermore, recall that the account vindicates and explains two core features of attention, viz. its ubiquity and heterogeneity. The former is explained by the entailment of attention by each mental act we perform, and the latter by the rich variety of things we do mentally. This richness causes attention to appear in various different guises, across different modalities, and in very different environments, and to be deployed for various different ends and with various different intentions in mind. Two other prominent accounts that are similarly animated by the twin features of ubiquity and heterogeneity – the selection for action view and attentional adverbialism – struggle to provide an adequate explanation.

Still, doubts may linger about the ultimate significance of attention on the proposed account. Can it make sense of the fact that attention matters to us? In fact, the account can readily explain our interest in attention. In the course of defending his account of knowing as the most general factive mental state, Williamson states that knowledge "matters to us because factive stative attitudes matter to us" (Williamson, 2000, p. 34). Paraphrasing his claim, we can say that attention matters to us because mental agency matters to us. To that, we may add that our

interest in *mental* agency, and hence our interest in when and whether attention is paid, is arguably due at least in part to the fact that mental agency is typically *covert*. To discover that a subject is paying attention is hence to discover that she is active when this fact may be harder to detect than in the more ordinary case, when her action is overt.[25,26]

## Notes

1. The debate over the scope of mental agency rages on (see, for example, Strawson, 2003; O'Brien & Soteriou, 2009; Upton & Brent, 2019; and several of the chapters in this volume). Consequently, with respect to some of the items listed in the text, there are those who will deny that they are indeed genuine instances of mental action. More on this below.
2. The heterogeneity of attention cannot of itself rule out the possibility that the different mechanisms or functional roles etc. are ultimately subsumable under one sufficiently general such mechanism or role. The present paper does not rule out this possibility, but it does suggest that an altogether different approach to explaining the nature of attention is more promising.
3. This section draws on (Levy, 2019a). See §3 and §4.2 of that paper for a more comprehensive discussion.
4. Cf. Bradley (1886, p. 316): "Any function whatever of the body or the mind will be active attention if it is prompted by an interest and brings about the result of our engrossment with its product. There is no primary act of attention, there is no specific act of attention, there is no one kind of act of attention at all." See also White (1964, pp. 5–8).
5. Mole may wish to reconcile his view with the existence of intrinsically attentional deeds. As Sebastian Watzl pointed out to me, Mole may attempt to do so by suggesting that such acts require cognitive unison. And since on his view, attention consists in cognitive unison, it would follow that the acts in question cannot be performed without attending. However, this only leads back to the other problem with the cognitive unison view noted in the text – namely, its implausible denial of partial or divided attention. It is deeply counterintuitive to suggest that we listen to, look at, or taste something *only if* all our cognitive resources are devoted to doing so, without even the slightest degree of attention paid at the same time to some other object. See the references in the text for discussion.
6. A more elaborate discussion of this objection, including possible replies on behalf of adverbialism, may be found in Levy (2019a, §3.2).
7. Watzl (2011) makes a similar point.
8. We can now see that Wu's example of kicking a ball, cited in the text, is in fact somewhat misleading as a representative illustration of how the selection for action view works. For the kicking is a *near-instantaneous* act – a fact that obscures the more extended temporal profile of many other ordinary acts.
9. A disjunctive account, claiming that attention consists in *either* solving the MMP or sustaining an existing solution might get around the objection in the text. But it would clearly drop points for lack of elegance and parsimony, and may be suspected of being ad hoc.
10. A somewhat different strategy for handling the objection from temporal mismatch may be suggested by Wu's talk in places of the 'non-deliberative' version of the MMP, which attention also solves. According to Wu, briefly,

the problem in this version arises because of "the demands of producing and guiding movement", and thus involves selection at a level considerably more fine-grained: "Ultimately, [solving the non-deliberative MMP] requires the constant and accurate representation of various magnitudes including the spatial location and dimensions of the target, as well as the speed and direction of movement" (Wu, 2008, p. 1007; see also Wu, this volume).

Now this sort of selection may perhaps occur throughout the entire course of acting. Yet this fine-grained selection is not plausibly a function attention performs at all. The reason is that the sort of magnitudes at play are (almost always) not represented in the contents of personal-level states and processes, but rather certain *sub*-personal states and processes. But attention, as nearly everyone recognizes, is a personal-level phenomenon. (Thanks to Wayne Wu for discussion here.)

11. For an argument that solving the MMP is not necessary for *action*, see Jennings and Nanay (2016). Wu (2019) replies to Jennings and Nanay.
12. The possibility of attending to just one object with no available alternatives, as in cases of degenerate selection, arguably impugns not just the 'selection for action' view but more broadly any view that subscribes to some version of the idea that attention is essentially *selective*. The point cannot be argued for here but if sound, it threatens to overturn what is seen by many as a truism – indeed, as the starting point for theorizing about attention (cf. for example the opening sentence of the entry on attention in the *Stanford Encyclopedia of Philosophy* [Mole, 2017]: 'Attention is involved in the selective directedness of our mental lives. The nature of this selectivity is one of the principal points of disagreement between the extant theories of attention.') For a different argument that attention does not consist in selection of any sort, see Levy (2019a, §4.2).
13. Cf. the Introduction to a recent volume on time and the philosophy of action, which treats the claim that action takes time as one of the "ordinary facts about action and agency" (Altshuler & Sigrist, 2016, p. 1). And similarly, the entry on "The Representation of Time in Agency" in a recent *Companion to the Philosophy of Time* sets out from the observation that "Our doings as agents in the world are irreducibly temporally extended" (Andersen, 2013, p. 470).
14. Notice that if the entire episode of (e.g.) calling some fact to mind occurs in the absence of any conscious awareness, with one being aware only of the moment when the information pops into one's head, this seems like a case where *attention* is altogether absent, too. Nevertheless, this does not make for a counterexample to *Entailment* since the process one is undergoing does not amount to action. To fully support this last point would require defending a substantive criterion of mental action. But its plausibility can be quickly verified by noting that, quite generally, one's lack of awareness deprives one of a standard agentive capacity – namely, the capacity to *control* what is taking place. More on this below. (Thanks to Sebastian Watzl for helpful comments here.)
15. In some cases, one's reasoning or deliberation will tend to be more open-ended, with at least some of the stages in the reasoning/deliberation process focused on a general question such as 'What shall I do?', or 'What to believe?', rather than on a specific act-type or proposition. In those cases, we can, in line with *Entailment*, see one's deliberation about e.g. what to do as involving attending to the question of what to do.
16. Cf. James (1890, p. 562): "Effort of attention is … the essential phenomenon of will", and James (1890, p. 424): "Volition is nothing but attention."

17. Another central and pertinent distinction between kinds of actions, besides the mental/bodily distinction, is the one between intentional and non-intentional actions. The account of attention being developed is meant to apply to both kinds. We return to this point below.
18. Famously, Williamson's epistemology stands in stark contrast to the traditional quest for a reductive analysis of knowledge. He rejects the attempt to decompose knowledge into supposedly more fundamental constituents, instead treating knowledge as a primitive. In drawing inspiration from Williamson, the present account of attention does *not* subscribe to any parallel ambitions; construing attention as a primitive plays no part in the argument of this paper (though the thought is not excluded, either). This implies that, for one thing, the account is entirely compatible with the availability of a sub-personal explanation of attending, of the sort pursued by psychologists.
19. Hyman proposes a further explanation for why knowing functions as a precondition for other factive states, building on his own account of knowledge. According to Hyman's explanation, "to know a fact is to be able to be guided by it, in other words, to be able to respond to it rationally, or take it into consideration or account" (Hyman, 2014, p. 565).
20. Watzl (2018) briefly touches on the possibility that 'the varieties of attention' should be understood as determinations of attention.
21. The result that attention is a determinable property of which all mental acts are determinates can be significant even independently of any attempt to explain the nature of attention. For example, the result can explain why, as I argue in Levy (2016, pp. 68–70), the "causalist" programme in philosophy of action struggles to accommodate instances of attending.
22. Compare Wu's discussion of attentional capture in Wu (2014, pp. 91–93), which the text draws on. And see also Watzl (2017, ch. 3).
23. No doubt stopping or refraining from attending can be *difficult* on occasion, e.g. when one is strongly curious about the context in which one's name is mentioned. But this does not provide a reason to doubt one's capacity to control the attentive stage: like many other powers and capacities we possess, this one too can sometimes be hard to exercise.
24. See also Koralus, 2014; Watzl, 2017; and the psychological theories mentioned in §2. While Adverbialism was rejected above, it bears an important structural similarity to *Determination* which should be noted: both views ultimately deny that attention to *V* is instantiated by some process independent of *V*-ing itself. Thanks to David Jenkins for discussion here.
25. Elsewhere I argue that the mental/bodily act distinction should in fact be supplanted by the covert/overt act distinction (Levy, 2019b). If the claim is sound, the argument of the present paper may need to be recast as supporting the idea that attention is the most general *covert* act-type.
26. For extremely helpful comments and discussion of material in this paper, many thanks to David Jenkins, Conor McHugh, Sebastian Watzl, Daniel Whiting, and Wayne Wu. The paper also benefited greatly from discussions at the University of Southampton, and the 2019 conference of the European Society of Philosophy and Psychology. I'm very grateful to the audiences there for their comments and questions.

## References

Allport, A. (1987). Selection for action. In H. Heuer, & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 395–419). Mahwah, NJ: Lawrence Erlbaum.

Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of 25 years. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 183–218). Cambridge, MA: MIT Press.

Altshuler, R., & Sigrist, M. J. (Eds.). (2016). *Time and the philosophy of action*. New York: Routledge.

Andersen, H. (2013). The representation of time in agency. In A. Bardon, & H. Dyke (Eds.), *A companion to philosophy of time* (pp. 470–485). Hoboken, NJ: Wiley-Blackwell.

Bradley, F. H. (1886). There any special activity of attention. *Mind*, *11*(43), 305–323.

Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon Press.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Hyman, J. (2014). The most general factive mental state. *Analysis*, *74*, 561–565.

James, W. (1890). *The principles of psychology*. New York: Henry Holt.

Jennings, C. D., & Nanay, B. (2016). Action without attention. *Analysis*, *76*, 29–36.

Koralus, P. (2014). The erotetic theory of attention: Questions, focus, and distraction. *Mind and Language*, *29*, 26–50.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*, 183–197.

Levy, Y. (2016). Action unified. *The Philosophical Quarterly*, *66*, 65–83.

Levy, Y. (2019a). Is attending a mental process. *Mind & Language*, *34*, 283–298.

Levy, Y. (2019b). What is 'mental action.' *Philosophical Psychology*, *32*, 971–993.

Malebranche, N. (1997). *Malebranche: Dialogues on metaphysics and on religion* (N. Jolley & D. Scott, Eds.). Cambridge: Cambridge University Press.

Mole, C. (2011). *Attention is cognitive unison*. Oxford: Oxford University Press.

Mole, C. (2017). Attention. In E.N. Zalta (Ed.). The Stanford Encyclopedia of Philosophy (Fall 2017 Edition). Retrieved from https://plato.stanford.edu/archives/fall2017/entries/attention/

Neumann, O. (1987). Beyond capacity: A functional view of attention. In H. Heuer, & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 361–394). Mahwah, NJ: Lawrence Erlbaum.

O'Brien, L., & Soteriou, M. (2009). *Mental actions*. Oxford: Oxford University Press.

Strawson, G. (2003). Mental ballistics or the involuntariness of spontaneity. *Proceedings of the Aristotelian Society*, *103*, 227–257.

Treisman, A. (1993). The perception of features and objects. In A. D. Baddeley, & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control* (pp. 5–35). Oxford: Clarendon.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Upton, C. L., & Brent, M. (2019). Meditation and the scope of mental action. *Philosophical Psychology*, *32*(1), 52–71.

Watzl, S. (2011). Review of "Attention is cognitive unison" by C. Mole. *Notre Dame Philosophical Reviews*. Retrieved from https://ndpr.nd.edu/reviews/attention-is-cognitive-unison-an-essay-in-philosophical-psychology/

Watzl, S. (2017). *Structuring mind*. Oxford: Oxford University Press.
Watzl, S. (2018). Review of "Attention, not self" by J. Ganeri. *Notre Dame Philosophical Reviews*. Retrieved from https://ndpr.nd.edu/reviews/attention-not-self/
White, A. R. (1964). *Attention*. Oxford: Blackwell.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Wu, W. (2008). Visual attention, conceptual content, and doing it right. *Mind*, *117*, 1003–1033.
Wu, W. (2011). Attention as selection for action. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 97–116). Oxford: Oxford University Press.
Wu, W. (2014). *Attention*. New York: Routledge.
Wu, W. (2019). Action always involves attention. *Analysis*, *79*(4), 693–703.

# 5 Mental Action and the Power of Effort

*Michael Brent*

## 5.1 Introduction

Much of what happens in consciousness seems to occur involuntarily. For instance, smelling a familiar aroma might trigger a fond memory, seeing your next-door neighbor might remind you of a neglected promise, and hearing a loud noise outside might result in your deciding to shut the window. In these cases, content is delivered to the mind involuntarily, triggered by the relevant stimuli in such a way that is not under your control.[1] However, not every case happens like this. In other situations, content comes to mind intentionally, where its delivery is under your control. For example, consider what happens when you are consciously deliberating about what to do, judging that something is the case, or calculating the tip at a restaurant. In such circumstances, content is delivered to the mind intentionally, in a way that you are making happen. How do you go about doing this?

Answering this question has not proven easy. On the standard accounts of intentional action, when your mental events appropriately cause and sustain the corresponding movements of your body, the result is an intentional action.[2] The standard accounts are reductive insofar as intentional actions are not metaphysically fundamental, but are reduced to appropriate causal relations between your mental events and the corresponding movements of your body. In the case of conscious mental action, the standard accounts run into a problem.[3] Consider what happens when you make a conscious choice or decision. Suppose while at the restaurant after consciously deliberating for a few moments you decide to eat salad rather than pizza. According to the standard accounts, if your decision to eat salad is a conscious mental action, then it must be the appropriate effect of a mental event, such as a desire or intention to make just that decision. But, if that is so, then the content of your decision is *already* present to mind as part of your desire or intention to make just that decision. This is problematic, for if the relevant content is already present to the mind, then that decision is explanatorily *redundant*. As Levy (2016, pp. 70–71) puts it, defenders of the standard

DOI: 10.4324/9780429022579-6

account "would postulate a decision to V caused by a desire or an intention to decide to V. That this is an incorrect analysis is easily seen from the fact that a decision to V is obviated by a prior intention to decide to V. If there is any recognizable sense in which one can intend or desire to decide to V, then one thereby makes V-ing part of one's plans, and hence in effect already decides to V." Hence, when applied to the case of conscious mental action, it seems the standard accounts are in trouble.

Could defenders of the standard accounts avoid this problem by claiming that the causally relevant mental events are not conscious? For instance, that an unconscious desire or intention to eat salad appropriately causes the delivery to consciousness of the relevant content? Unfortunately, this move does not help. On the standard accounts, the causally relevant mental events rationalize the actions they cause – they explain why the agent did what they did, by showing what it was about the action that appealed *to the agent*. If the causally relevant mental events and their content are not conscious, then we sever this connection with the agent. But, if the agent is not aware or conscious of the causally relevant mental events and their content, the agent cannot explain why they made the decision that they did, and so it fails to be intentional on their part. This is problematic, for conscious decisions are among those aspects of our mental lives for which we have rationalizing explanations. More importantly for present purposes, this would leave unexplained how content is intentionally delivered to consciousness in the first place.

Likewise, could a proponent of the standard account avoid this problem by claiming that the content of the causally relevant mental events differs from the content of the resulting state of mind? Here, too, this move will not work. On the standard accounts, if the content of the causally relevant mental events does not specify the precise action that you desire or intend to perform, it is no longer *intentional* on your part.[4] For instance, if the content of the relevant mental event is to reach *some* decision about what to eat, this is not specific enough to render the subsequent decision to eat salad intentional on your part. The content of this mental event is not precise enough to bring to mind the particular decision to eat salad, rather than pizza. Compare this case to a bodily action. If you desire or intend to make *some* bodily movement, this is not precise enough to render the subsequent upward movement of your left arm intentional. For the content of such a desire or intention is not precise enough to bring about that particular upward movement of your left arm, rather than another movement. In both cases, we are no longer in a position to explain the precise decision or bodily movement that occurred, as opposed to another, and so neither are intentional on your part.

In what follows, my main goal is to introduce an alternative account of conscious mental action.[5] On the account presented here, conscious mental action is not explained reductively in terms of appropriate causal relations between mental events. Rather, I suggest that conscious mental

actions necessarily involve an agent *doing something* to bring about an effect. More specifically, I claim that when you are performing a conscious mental action, you are playing a necessary and ineliminable causal role in the process by which you produce and sustain it. I explain your causal role in terms of your cognitive capacities, such as memory, imagination, attention, and so forth, and argue that you use such capacities by doing something in particular – namely, by exerting effort.

Several assumptions are worth making explicit before we proceed. Unconscious mental action will not be discussed here, and neither will the question of doxastic freedom. So, for example, whether judgement and choice are autonomous, or whether belief-acquisition is the involuntary product of the operation of cognitive capacities, is not discussed.[6] In addition, I will not address the question of what makes a particular conscious mental action an instance of, say, good or bad reasoning.[7] And, I make no assumptions about the ontological status of the mental. I will, however, assume that you can be aware of *how* content is delivered to consciousness – e.g., whether intentionally or not.[8]

With these caveats in mind, the paper proceeds as follows. I begin with Galen Strawson's well-known skeptical account of conscious mental action (Section 5.1). Beginning with Strawson is helpful, in part because his skeptical account is offered in light of the above problem faced by standard accounts, as well as recent findings in psychology suggesting that much of what happens in consciousness is involuntary.[9] In light of both, Strawson claims that mental action is restricted to the act of triggering a ballistic process that automatically results in the delivery of content to consciousness. I argue that although his account contains an important insight, there are mental actions that are not accommodated by Strawson's skeptical framework, so it falls short (Section 5.3). Then, I outline an alternative account of mental action described in terms of the cognitive capacities that you employ by exerting effort (Section 5.3). Since the alternative presented here is a first attempt at shedding light on a new explanatory framework, I end by briefly noting some of the work that remains (Section 5.4).

## 5.2 Going ballistic

Galen Strawson claims that much of what takes place in consciousness is involuntary and merely reflexive.[10] He suggests that although mental action exists, it is restricted to the act of triggering an event in which content is delivered to consciousness. The idea is that the delivery of content is not a mental action, or any part of one, since the content in question has not been intentionally assembled or selected as such. Rather, for Strawson, mental action is restricted to the act of triggering a ballistic process that automatically results in the delivery of content to consciousness. The act of triggering this ballistic process can take a variety of

forms, but in each case, your control is limited to triggering the relevant mental event. Once you have triggered the relevant event, you no longer control what is happening. He summarizes the claim as follows:

> No actual natural thinking of a thought, no actual having of a particular thought-content, is ever itself an action. Mental action in thought is restricted to the fostering of conditions hospitable to contents' coming to mind. The coming to mind itself – the actual occurrence of thoughts, conscious or non-conscious – is not a matter of action.
>
> (Strawson, 2003, p. 234)

Thus, for Strawson, mental action is limited to the act of triggering the relevant mental event – the rest is waiting, however briefly, for content to be delivered to consciousness.

Strawson is correct to say that you can trigger a mental event in which content is delivered to consciousness as a result, and that in such cases this content has not been intentionally constructed by that act of triggering. Indeed, it seems that in many cases, when you entertain content in thought, the relevant content is simply delivered to consciousness as such. The coming-to-mind of that content can be the effect of your mental action, what Strawson describes in terms of triggering. But, the specific content that comes to mind is not under your control. Why, though, should we limit our account of conscious mental action to the act of triggering the delivery of content to consciousness?

When describing his preferred notion of intentional action, Strawson cites the work of Donald Davidson, who once claimed that in the case of bodily action, you do nothing more than move your body – the rest is up to nature.[11] On this account, intentional action consists of just those bodily movements that you are able to bring about and control *directly*, without the use of causal intermediaries.[12] Given this account of intentional action, applied to the mental domain, your control over what happens in consciousness extends no further than what you can bring about directly without causal intermediaries. For Strawson, this is the act of triggering a mental event. After your act of triggering has occurred, the rest is a matter of ballistics, i.e., it is not under your control.

Unfortunately, Strawson has overlooked other types of mental actions, especially those that do not involve triggering a mental event in which content is delivered to consciousness. In the next section, I provide several examples of mental actions that are not limited to the act of triggering the delivery of content to consciousness, and so cannot be accommodated within Strawson's skeptical framework. I assume that variations of each example exist and occur quite frequently, as anyone who has undergone the relevant experiences can attest.

## 5.3 Beyond ballistics

In this section, I provide examples of mental actions that involve control over the *content* that is delivered to consciousness, and the *cognitive capacities* that are used when doing so. The examples demonstrate the wider extent of our powers of agency within the domain of the mental, and show that Strawson's skeptical account falls short in at least two ways: first, by overlooking mental actions that are not restricted to the act of triggering the delivery of content to consciousness, and second, by failing to explain how you go about performing intentional mental actions in the first place. The examples demonstrate that you can intentionally *remove* content from consciousness, *manipulate* features of content while it remains present to consciousness, and *ignore* irrelevant or disruptive content while remaining focused on the task at hand. Moreover, each example highlights the fact that, even in the case of triggering the delivery of content to consciousness, the agent in question *is doing something* when using the relevant cognitive capacities, something that is not accounted for by Strawson. Exactly what they are doing shall be addressed in the subsequent section (Section 5.3).

Consider first a situation in which you are in the midst of mind-wandering or daydreaming, and your attention has lapsed from the task at hand. Suppose that you then actively stop the daydream from occurring by suppressing the activity of the relevant cognitive capacity and re-focusing your attention on the task that you were performing prior to the onset of the daydream. In such cases, you intervene by doing something so as to take control of the relevant cognitive capacity – it takes some effort on your part to stop the occurrence of the daydream and to redirect your attention back to the original task. Arguably, the suppression of visual or auditory images is a clear case where you do not trigger an event in which new content is delivered to consciousness.[13] Rather, precisely the opposite takes place: by interfering with and stopping the activity of the relevant cognitive capacity, you remove content from consciousness and then bring your wandering mind back to the task at hand.[14]

Next, consider someone who is asked to visually imagine a particular object and then manipulate the various qualitative features of the image. Keeping the image present to consciousness requires that she continuously employ the relevant cognitive capacity as she attends to and intentionally manipulates the qualitative features of the image. Arguably, such transitions in imagined content do not involve bringing new content to consciousness. Rather, the same image is changed as she manipulates its qualitative features, such as when you imagine the visual appearance of a three-dimensional shape and rotate it mentally.[15] Such feats of visual imagination require the exertion of effort of your part, as anyone who has done so can attest, and when they occur it is arguable that the same

visual image present to consciousness over an extended period of time undergoes an intentional change in its imagined properties and spatial orientation. Crucially, it is by exerting effort and using the relevant cognitive capacity in a specific way, rather than triggering the delivery of new content to consciousness, that this change is intentionally produced.

Third, consider the extraordinary case of Cathy Hutchinson.[16] In 1995, Hutchinson suffered a catastrophic brainstem stroke that left her a tetraplegic. Ten years later, a microelectrode array was implanted in her primary motor cortex. After recovering from the surgery, Hutchinson began participating in weekly trials in which her neuronal activity was recorded during a variety of tasks, with the initial goal of developing and improving the brain-computer interface. The mechanism detects the activity of multiple neurons in the primary motor cortex, translates the patterns of neuronal activity into motor commands, and through a computer, controls another device in light of the cortical activity. Through repeated rehearsal, Hutchinson began controlling the movement of a cursor on a computer screen, much like you might do with a computer mouse. She was able to do this by imagining making the hand, wrist, and arm movements that she would have performed had she had the ability to use the relevant bodily capacities. Crucially, she managed to do so by controlling the bodily movements that she imagined herself performing, as well as fine-tuning what she was doing in response to her visual perception of the cursor's movements on the screen. Here, too, when imagining the relevant movements of her body, manipulating and controlling those imaginary movements does not deliver new content to consciousness. Rather, the same motor image that is present to consciousness over time undergoes an intentional change in its imagined features and spatial orientation.[17]

Finally, consider the practices of mindfulness and meditation. A growing body of empirical evidence indicates that such practices enable us to develop and perform mental actions that do not involve triggering the delivery of content to consciousness.[18] Studies have shown that mindfulness training improves performance on tasks where you attend to relevant information while suppressing irrelevant information, or overcoming a conflict caused by interfering information.[19] For example, when the content "I should check my email" suddenly arises unbidden in consciousness, instead of immediately pursuing this course of action or bringing to consciousness associated content (e.g., about impending deadlines, or neglected students), you can remain aware of that content while allowing it to pass from consciousness, as it were. In order to ignore such irrelevant content or overcome potentially disruptive content, you must exert a distinctive kind of effort to maintain control over the pertinent cognitive capacity. Through practicing mindfulness and meditation, performance on such cognitively demanding tasks can be improved.[20]

The point of these examples is twofold. First, they reveal the wider extent of your control within the domain of the mental by illustrating that there exist conscious mental actions that involve doing something besides triggering the delivery of content to consciousness. In addition, mental action also involves removing content from consciousness, changing the qualitative features of content that are present to consciousness, and ignoring irrelevant or potentially disruptive content, each of which requires effort and some degree of skill on your part. Second, they support the claim that, even in the case of triggering the delivery of content to consciousness, conscious mental action requires that you are *doing something* in order to perform the action in question. As I claim in the next section, when performing conscious mental actions, you play a necessary and ineliminable causal role in the process by which you produce and sustain it. So long as your causal role remains unexplained, not only do we fail to capture the full range of our powers of agency, we fail to explain how content is delivered to consciousness intentionally in the first place.

## 5.4 An alternative account of mental action

Thus far, I have suggested that the standard accounts of action face a significant challenge when applied to the mental, and that Strawson's ballistic account of conscious mental action falls short. By providing examples of such actions that are not restricted to triggering an event in which content is delivered to consciousness, I have claimed that there are types of mental actions not explained by his skeptical framework. In this section, I present an alternative account of conscious mental action described in terms of an agent exerting effort when using the relevant cognitive capacities. On the account presented here, conscious mental action is not explained in reductive causal terms. Rather, I will claim that it is only by doing something – namely, by exerting effort – that conscious agents succeed in performing such mental actions.

To begin, note that other than listing examples, Strawson says nothing about *how* you trigger the delivery of content to consciousness in the first place.[21] He hints that the cause of your doing so might be an event within your brain that happens outside the scope of consciousness, but says nothing other than suggesting that this might be the case.[22] However, unless we explain how you trigger the delivery of content to consciousness, the contrast introduced at the outset of this paper, between cases where content is delivered to consciousness involuntarily and cases where it is delivered intentionally, remains unaccounted for. As a result, we have yet to explain how you perform conscious mental actions in the first place.

This is the mental counterpart of a well-known problem in the case of bodily action.[23] The problem is that ballistic accounts of action cannot

differentiate between intentional bodily actions and mere bodily movements while they are happening, for after the triggering occurs the same kind of thing happens in both cases. In both cases, immediately following the triggering impulse, your bodily movements are occurring in a merely ballistic manner that is no longer under your control. As a result, this makes it impossible to explain the fact that while performing an action, you are causally related to your bodily movements in a way that is necessarily absent in cases where your bodily movements are merely happening. During an action, you are *controlling* the relevant bodily capacities, whereas in the case of mere bodily movements you are not. By restricting the scope of such control to the act of triggering an event, ballistic accounts of action cannot differentiate between these distinct ways in which you are related to your bodily movements while they are happening, and so they fail to explain how you perform bodily actions in the first place.

The analogous problem in the case of mental action is that ballistic accounts cannot differentiate involuntary content delivery and intentional content delivery while each is taking place, for after the act of triggering the same kind of thing occurs in both scenarios. In both, immediately following the act of triggering, the subsequent activity of your cognitive capacities is not under your control, and content is delivered to consciousness as a mere effect. As a result, this makes it impossible to explain the fact that while you are performing a mental action, you are causally related to your cognitive capacities in a way that is necessarily absent in cases where the activity of your cognitive capacities occurs involuntarily. During a mental action, you are *controlling* the relevant cognitive capacities, whereas when content is delivered to consciousness involuntarily you are not. By limiting the scope of your control to the act of triggering content delivery, ballistic accounts of mental action cannot differentiate between these different ways in which you are related to your cognitive capacities, and so they fail to explain how you perform conscious mental actions in the first place.

Recall that the standard accounts of action face a similar problem when applied to the mental. On such accounts, as a conscious agent, your causal role when performing mental actions is explained wholly in terms of causal relations among your mental events. Because such accounts assume that the relevant content is present to consciousness within the mental events that allegedly cause and sustain the resulting mental action, they take for granted the very thing that needs to be explained – *how* you intentionally deliver content to consciousness in the first place. In fact, I think this problem is symptomatic of a deeper metaphysical worry faced by standard accounts of action. On the standard accounts, the delivery of content to consciousness is appropriately caused by nothing but your conscious mental *events*. This is the core reductive commitment of such accounts, where conscious mental action

is reduced to and explained just in terms of appropriate causal relations among the relevant events. The deeper metaphysical problem concerns this reductive commitment.[24] If the delivery of content to consciousness is appropriately caused by nothing but your mental events, it follows that you are playing no causal role when content is delivered to consciousness. Your mental events transpire as a causal consequence of nothing but the occurrence of other events, so that as a conscious agent you are *not doing anything* to make such events happen. Hence, the standard accounts depict you as an inactive onlooker, someone who is merely aware of your own conscious mental events as they transpire, without intervening causally in the processes by which they are taking place. Since those mental events are caused by other events and not by anything that you are doing in order to make this happen, there is no sense in which you are doing anything here at all, and so there is, in fact, no intentional action in view.[25] As a result, the standard accounts fail to explain conscious mental action in the first place.

Thus, neither Strawson's ballistic account nor the standard reductive accounts explain how you perform conscious mental actions. Strawson assumes that you can trigger the delivery of content to consciousness, but does not say *how* you go about doing so. The standard reductive accounts assume that the delivery of content to consciousness is appropriately caused by nothing but your mental events, so that as a conscious agent you are not doing anything to make such events happen, and there is no intentional *action* in view here. In contrast to both, I suggest that the best explanation of how you perform conscious mental actions is in non-reductive causal terms. More precisely, I suggest that mental actions occur as a result of something that you are doing, something that is absent in cases in which content is delivered to consciousness involuntarily. Only by recognizing your necessary and ineliminable causal contribution to the etiology of mental action can we explain the different ways in which content is delivered to consciousness, and thereby explain what is missing in Strawson's ballistic account and the standard reductive accounts – your causal role as a conscious agent.

To account for your causal role when performing conscious mental actions, we should adopt a different framework for understanding the metaphysics of causation, one in which causation is not reduced to an asymmetric relation between events but is understood in terms of concrete particulars and the causal powers that they possess.[26] On such views, causation is not a relation among events, but is understood in terms of manifesting causal powers. When a physical object is manifesting its causal powers, it is necessarily causing an effect. Hence, on such views of causation, there is a cause, an effect, and a causing. Insofar as the causal powers that an object is manifesting are among its properties, the object is the *cause*; the manifesting of the relevant causal powers is *causing* the result in question, and the *effect* is what results from this

process.[27] Here, an event takes place when an object is manifesting its causal powers, but it is the object that produces the relevant effect by manifesting its causal powers, rather than the event in which this occurs.

Obviously, a full-fledged defense of this way of understanding the metaphysics of causation must be provided elsewhere. However, for present purposes, this sketch provides the resources for explaining your causal role when performing conscious mental actions, resources that are missing in Strawson's ballistic account, and the standard reductive accounts. As a conscious agent, you are a concrete particular that possesses numerous parts and properties, stands in various relations to other physical objects, occupies an area of space, and persists through time while undergoing change.[28] Among the properties that you possess are your cognitive capacities, such as imagination, memory, attention, and so forth. Whenever such capacities are manifesting, an effect is necessarily produced. In every case, the manifesting of a cognitive capacity delivers content to consciousness, in a capacity-specific way. For instance, when your capacity for episodic memory is manifesting, content is delivered to consciousness as remembered, and when your capacity for visual imagination is manifesting, content is delivered to consciousness as visually imagined. Crucially, the manifesting of a cognitive capacity can be intentional on your part, or it can be involuntarily triggered by other stimuli. The difference depends in part upon *how* the cognitive capacity is manifesting on that occasion.

I suggest that as a conscious agent, when performing a mental action, you are manifesting the relevant cognitive capacity by doing something. Specifically, you are exerting effort in the process of manifesting the relevant cognitive capacity and causing the delivery of content to consciousness. Discussion of effort remains largely absent from recent philosophical literature, even though it is a pervasive feature of our lives.[29] In the mental domain, effort is often described in terms of cognitively demanding actions, such as the examples discussed above (Section 5.2).[30] By contrast, on the alternative account introduced here, exerting effort is not restricted to the performance of conscious mental actions that are difficult or demanding. Rather, exerting effort is understood more broadly than this, to include not only the performance of cognitively demanding tasks, but every conscious mental action that you perform intentionally.[31]

Effort has several key features that are worth highlighting. On the alternative account introduced here, effort is a causal power that conscious agents employ directly as such, so that exerting effort *just is* manifesting a causal power they possess.[32] Whenever causal powers are manifesting, they make something happen, e.g., they produce change.[33] In the case of conscious mental actions, exerting effort occurs together with the manifesting of the relevant cognitive capacities. Your cognitive capacities are causal powers too, but when manifesting they deliver

content to consciousness, whether intentionally or otherwise. Exerting effort is *how* you wield causal control over your cognitive capacities when delivering content to consciousness intentionally. Effort is thus a causal power through which you express your intentional agency when using your cognitive capacities to perform conscious mental actions.[34] Contrast this with the standard accounts of action, where intentional agency is expressed through the causal efficacy of those mental events that explain why you perform that action. On the alternative introduced here, intentional agency is given expression by something that you are doing – exerting effort – through which you are producing the relevant conscious mental actions as a result.

Given that exerting effort is manifesting a causal power, doing so always occurs together with the manifesting of other capacities, jointly causing something to happen. As a result, in no situation are you exerting effort *without* producing an effect of some kind.[35] In typical conditions where all goes well, exerting effort occurs simultaneously with the manifesting of the relevant cognitive capacity, together causing the delivery to consciousness of the sought-for content, as well as the sorts of actions discussed above (Section 5.2). Crucially, whether the *relevant* cognitive capacity is manifesting on any given occasion is determined by the content of the beliefs, desires, intentions, and other states of mind that explain why you are performing the action in question. For instance, if you want to remember what you ate yesterday for lunch, the content of your desire determines which specific cognitive capacity is relevant on that occasion. It specifies that the relevant cognitive capacity is episodic memory, rather than imagination, say. Equally as important, whether exerting effort and the relevant cognitive capacity *successfully* produce an intentional action is partly determined by the content of such states of mind, and also partly determined by the type of mental action in question, the features of the agent, and the wider context. For instance, if you ask me to read and understand a text written in a language that I do not at present comprehend, I will fail to perform that mental action (assuming that reading is such an action). Because I lack the relevant cognitive capacities, any exertion of effort on my part will fail to produce the sought-for result. In such atypical conditions where all does not go well, an intentional action is not performed, either because content that was not sought-for is delivered to consciousness, or the relevant cognitive capacity is not employed. Exerting effort is thus *disjunctive*: either you are exerting effort together with manifesting the relevant cognitive capacity when acting intentionally, or you are exerting effort together with manifesting another cognitive capacity, thereby producing a different kind of result.[36] Again, in no case are you exerting effort and producing nothing as a result.

Note that when you succeed in performing a conscious mental action, exerting effort while manifesting the relevant cognitive capacity is

causally basic on your part.[37] It is causally basic because there is nothing else that you do, and no other action that you perform, by the doing of which, that you are causing the delivery of content to consciousness. Rather, exerting effort while manifesting your cognitive capacities *just is* your causing of that effect – that is, delivering content to consciousness *is* the causally basic mental action that you are performing, which is explained in terms of a concrete particular manifesting its causal powers. And, although doing so is causally basic on your part, it does not follow that it is spontaneous, in the sense that there is no accounting for *why* you perform the intentional mental actions that you do.

For instance, suppose I ask you to recall specific information, you agree to my request, and then go about doing just that. On the alternative account introduced here, as an intentional mental action that you are performing, you are delivering content to consciousness by exerting effort and manifesting the relevant cognitive capacities. You do so in part because of my request, but since the delivery of that content results from your doing something – namely, exerting effort and manifesting the relevant cognitive capacities – it is intentional on your part. Compare this to a slightly different scenario in which upon hearing my request to recall that information, your comprehending perception of what I say immediately triggers the delivery of that content to consciousness. Here, the relevant content is delivered to consciousness, but not as a result of your doing something in order to make this happen. The delivery of that content is not intentional on your part, so what takes place is an involuntary mental event, rather than an intentional mental action. On the alternative account introduced here, your beliefs, desires, intentions, and other mental events explain why you perform the mental actions that you do, but their occurrence does not cause the action in question.[38] Because these mental events do not cause the intentional mental actions you perform, there is a need for another explanation of the process by which you cause and sustain those actions. According to the alternative account presented here, that causal role is occupied by your doing something – exerting effort in the process of using the relevant cognitive capacities. Thus, in addition to those mental events that explain why you act as you do, exerting effort is a necessary feature of the process by which you are acting, where your basic mental action *just is* your delivering the relevant content to consciousness by that means.

Note, too, that the alternative introduced here allows us to explain the fact that we typically hold each other *accountable* for the conscious mental actions that we perform, a crucial aspect of our lives that Strawson's ballistic account and the standard reductive accounts have trouble explaining, since neither accounts for your necessary causal role as conscious agent. This point is especially salient during conscious mental actions that develop over time, such as episodes of practical reasoning or deliberation. We hold each other accountable for performing such actions.

We praise each other for deliberating well, and criticize each other for deliberating poorly. We admire those who think carefully and consider as many options as feasible, and we admonish those who rush to hasty conclusions. Arguably, these practices require that conscious agents persist throughout the mental actions that they perform, for it is the deliberator herself that we praise and admire, not her conscious states of mind and the causal relations between them. For this practice to be intelligible, she must persist throughout the duration of her deliberations, however brief.

To appreciate why conscious agents must persist throughout their mental actions, suppose that as a conscious agent you do *not* persist throughout the duration of the mental actions that you perform, but exist as a series of discrete conscious mental events, one after another, standing in various relations to each other.[39] Suppose, too, that such mental events are individuated in part by the content that is present to consciousness at that time. The problem is that when you are engaged in a temporally-extended conscious mental action like practical deliberation, the content of the relevant mental events stand in various relations to one another, relations that, while deliberating, you must in some way be aware of as such.[40] For instance, when deliberating about whether to eat salad or pizza for lunch, in thinking (1) that the salad is the healthiest option and (2) that you value a healthy diet and lifestyle, when you conclude (3) that you will eat salad rather than pizza, this requires that (3) is concluded *because* of its relation to (2) and (1). In turn, this requires that there exists something that integrates the content of, and relations between, these conscious mental events across time, so that together they serve as the basis upon which you draw your conclusion.

I suggest that both requirements necessitate that there is a single, numerically distinct conscious agent that persists throughout the duration of these mental events, through whose ongoing existence as such the content of those events is integrated over time, on the basis of which the conclusion is drawn. That is, in concluding that (3), you are aware that that content is delivered to consciousness precisely as the conclusion of that episode of practical deliberation, and depends upon the relations between (1) and (2). It is precisely because you are aware that the salad is the healthiest option, and that you value a healthy diet and lifestyle, that you then conclude as you do. Such episodes of practical deliberation require that as a conscious agent you persist continuously throughout this temporally-extended conscious mental action as one, numerically distinct conscious agent, aware of the various relations between the content of those mental events, and integrating them together within this action that develops over time. This is not compatible with the claim that your existence as a conscious agent is reduced to, or consists in, a series of distinct conscious mental events.

On the alternative account of conscious mental action offered here, this problem does not arise. As a conscious agent, you are a concrete

particular that persists through various sorts of change, and has numerous parts and properties, including the relevant cognitive capacities that you employ during the performance of conscious mental actions.[41] As such, you are not identical with any particular conscious mental event, or any collection of such mental events taken together. If you were identical with a particular conscious mental event, then your existence would last only so long as that event; and, if you were identical with a collection of mental events taken together, then there would exist as many of you as there were conscious mental events. Neither option is plausible on metaphysical grounds. On the alternative view presented here, you are not identical with any conscious mental events, individually or collectively, and you are not ontologically distinct from what takes place in consciousness, something that would exist in the absence of all conscious mental actions that you perform or conscious mental events that you undergo as subject. Rather, as a concrete particular, you are the conscious agent for whom it is like something to experience conscious mental events and to perform conscious mental actions.

## 5.5 Conclusion

I have suggested that the standard accounts of action face a significant challenge when applied to the mental, and that Strawson's ballistic account of mental action falls short. Although his account contains an important insight, there are types of mental actions that are not restricted to the act of triggering the delivery of content to consciousness. On the alternative presented here, it is by exerting effort that you control your cognitive capacities and deliver content to consciousness during a mental action. The notion of effort was explained in terms of a causal power that you possess as a concrete particular. It is by exerting effort and manifesting your cognitive capacities during the performance of a mental action that you are delivering content to consciousness as a result.

The account of conscious mental action presented here is a first attempt at shedding light on an alternative explanatory framework. Providing a comprehensive defense of this framework requires additional work. For instance, the account presented here has assumed that as a conscious agent you are a concrete particular or physical object. A complete account of conscious mental action must justify this key background assumption. In addition, the constitutive conditions under which exerting effort successfully combines with the relevant cognitive capacities must be spelled out, and, although not explored here, there is an important epistemic component of mental action that must be explained as well. Conscious agents who possess the relevant epistemic capacities can come to know which mental phenomena have been brought about and controlled by themselves and which have not. A complete account of conscious mental action must explain the ways in which such agents are able to achieve

this kind of self-knowledge. Furthermore, little has been said about the rational and normative dimension of mental action, which is of crucial importance. Agents who possess the relevant rational and normative capacities will be in a position to assess the reasons that support their performance of a particular conscious mental action. A full-fledged account of such action must explain the ways in which we perform mental actions in light of our assessments of the relevant reasons. Last, but not least, the potential implication that the account has for the issue of doxastic freedom is worthy of exploration. Such matters await further investigation.[42]

## Notes

1 The notion of content here is conscious representational content. It is neutral with regard to whether it is conceptual or non-conceptual, and to whether its structure is Russellian or Fregean or otherwise. Consciousness is here what Block (1995) describes as phenomenal consciousness, which consists in there being some way it is like for you to be in that state of mind. This, of course, derives from Nagel (1974). Content delivery is the idea that we might express by saying that content "comes to mind", or becomes an "object" of conscious awareness or attention.

2 For defense of standard accounts, see, for instance, Bishop (1989), Brand (1984), Bratman (1987), Davidson (1963), Enç (2003), Goldman (1970), and Mele (1992). The qualifier "appropriate" rules out deviant causation, and "mental event" is used broadly to encompass mental states and processes.

3 For discussion of a similar problem, see Levy (2016), Mele (1992, 1997), and Wu (2013).

4 For discussion, see Shepherd (2015).

5 Until recently mental action received scant attention in philosophy of mind and action. Exceptions include Boyle (2009; 2011), Buckareff (2005; 2007), Hunter (2003), Levy (2016), Mele (1997; 2009), O'Brien and Soteriou (2009), Peacocke (2007), Proust (2001; 2013), Ruben (1995), Shepherd (2015), Soteriou (2013), Strawson (2003), Valaris (2016), Watzl (2017), and Wu (2013).

6 For discussion see, for instance, Arpaly and Schroeder (2012; 2014), Boyle (2011), Chrisman (2016), Hieronymi (2009), and Paul (2015).

7 See Boghossian (2008), Broome (2013), Harman (1999), McHugh and Way (2018), and Wedgwood (2006), among others, for discussion.

8 For defense of a role of awareness in supporting introspection, see Watzl (2017, chap. 11). There is growing evidence suggesting that introspection can be improved with practice. For instance, see Baird, Mrazek, Phillips, and Schoolar (2014), Hill and Updegraff (2012), and Hurlburt and Heavey (2001; 2004). For an overview, see Schwitzgebel (2016).

9 The source of such research can be traced to Kahneman (1973) and Schneider and Shiffrin (1977). More recent work includes Bargh (1994) and Bargh and Chartrand (1999). See Evans (2010) for a useful review.

10 Strawson (2003). See also Wu (2013) for a similar view. Strawson (2003, p. 246, n. 41) cites Libet (1985; 1987; 1989), Wegner (2002), and Wegner and Wheatley (1999) for empirical evidence in support of his claim about mental action. For criticism of the relevant empirical literature, see Bayne (2006) and Mele (2009).

11 See Davidson (1971). It is worth noting that Davidson eventually changed his mind. See Davidson (1978). When Strawson (2003, p. 245) quotes Davidson, he says: "In cognition we never do more than aim or tilt our minds; the rest is up to nature, trained or not. Much bodily movement is ballistic, relative to the initiating impulse; the same goes for thought."
12 This is why Strawson (2003, p. 245, n. 39) claims that we could distinguish between intentional actions and what merely happens more radically than Davidson (1971), e.g., inside the brain. If this were true, *all* intentional action would be nothing but triggering events in the brain. The similarity to a view of agent causation once defended by Roderick Chisholm (1964/2003) is striking.
13 Strawson might admit that, strictly speaking, although no new content is delivered to consciousness, this is an instance of triggering a ballistic process of, say, "clearing your mind." See, e.g., Strawson (2003, p. 232), where he seems to admit this. Not only would this involve a change in his official view, Strawson would owe an account of *how* you trigger the delivery or removal of content when you do, which he does not here provide.
14 See Smallwood and Schooler (2015) for a review of empirical literature on mind-wandering.
15 See Shepard and Metzler (1971) for a classic experiment involving mental rotation of images. For an overview, see Nigel (2017).
16 Her case is described in Hochberg et al (2012). A similarly incredible case is described in Bouton et al (2016).
17 Note the similarity with the previous case, except here we need not assume that Cathy Hutchinson is rotating a visual image mentally. It seems best to think of what she is doing as imagining the movements of her body, i.e., rotating a motor image mentally. For related discussion of motor imagery, see Mast, Bamert, and Newby (2007).
18 See Upton and Brent (2019) for discussion.
19 For discussion of the empirical literature, see Li, Liu, Zhang, Liu, and Wei (2018).
20 In such cases, it's possible that with practice the degree to which one must exert effort diminishes over time.
21 See Strawson (2003, pp. 231–232).
22 See Strawson (2003, p. 248, n. 45) where he suggests that "the natural causality of the whole engine of innate mental equipment [is] activated and tuned by experience" in a way that is "automatic and standardly involuntary." He does not provide an argument in defense of this controversial remark, and, as mentioned in note 10, the empirical work that he cites has been widely criticized.
23 For discussion of the problem in the context of bodily action, see Frankfurt (1978).
24 In fact, the metaphysical problem about to be raised does not require that mental *events* play the relevant causal role. The problem arises so long as any part or property of you, or system within you, occupies that causal role on your behalf, as it were. This should become clear in the main text.
25 This echoes another well-known problem in the case of bodily action, that of the disappearing agent. See Hornsby (2004), Shoemaker (1988), and Velleman (1992) for discussion.
26 For defense of causal powers-based accounts of causation, see Bird (2007), Ellis (2001), Harré and Madden (1975), Heil (2012), Lowe (2008), Martin (2008), Molnar (2003), Mumford and Anjum (2011), and Whittle (2016). There are differences between these views, most important of which is the role of particular substances in causation. Here, I assume that causal

powers are properties had by particular substances, where substances are persisting physical objects, rather than bundles of properties. For an account of substance, see Hoffman and Rosencrantz (1994).

27 Note that there are cases in which an effect might occur simultaneously with the manifesting of the relevant causal powers, so the effect need not be temporally distinct from the process that is causing it. Thanks to Lisa Miracchi for suggesting that I clarify this here.

28 Though I assume that as a conscious agent you are a concrete particular (i.e., a persisting physical object), the account presented here does not require this assumption. Exactly what kind of concrete particular (e.g., a person, an animal, a brain, etc.), I set aside for present purposes.

29 For some exceptions, see Bradford (2015), Brent (2017), Holton (2009), Kane (1996), and Shepherd (2016).

30 For instance, see Hagger, Wood, Stiff, and Chatzisarantis (2010), who identify effort with the depletion of a limited cognitive capacity, or Kurzban, Duckworth, Kable, and Myers (2013) and Shenhav et al (2017), who hold that effort is the experiential counterpart of opportunity-cost calculations that play a role in cognitive-control allocation processes. See, also, Friese, Loschelder, Gieseler, Frankenbach, and Inzlicht (2018) for recent criticism.

31 Note that effort is not limited to the mental domain. For an account of the causal role of effort in bodily action, see Brent (2017). An implication here is that there is no effortless intentional action. Actions we might describe as effortless are such that when performing them, you are not aware of yourself *as* exerting effort.

32 See Diamond (2013) for an overview of such executive-level capacities.

33 Note that when manifesting, causal powers make something happen, but not every causal power thereby produces change. When manifesting, some causal powers *maintain* the ongoing existence of a state of affairs. For discussion of this point, see Williams (2014; 2017).

34 On the account introduced here, the relation between the causal power you manifest when exerting effort, and the cognitive capacities used during conscious mental actions, is partly constitutive of intentional mental agency. The relation is not itself directly controlled by the agent. Rather, it is among the conditions that make such agency possible in the first place. Such conditions must be discussed in future work.

35 Could one exert effort *without* manifesting any cognitive capacity whatsoever? For instance, when you try to remember a name but nothing comes to mind, have you exerted effort without causing any result? On the alternative presented here, something happens in conjunction with your exerting effort, though not the action that you seek to perform. In the case of trying to remember a name, your capacity for semantic memory is activated as you attempt to recall the information, even though doing so fails to deliver content.

36 For discussion and defense of disjunctive accounts of action, see Hornsby (2008) and Ruben (2008).

37 See Hornsby (2013) for defense of a notion of basic activity.

38 See Sehon (2005; 2016, chap. 2) for defence of a non-causal, teleological account of the role of desires, beliefs, intentions, and other mental events in the explanation of intentional mental action. See also Anscombe (1957). I set aside exploration of this for another occasion.

39 The sort of view I have in mind here is exemplified by broadly "Lockean" accounts of diachronic personal identity. Roughly put, such a view holds that, as a conscious agent, your existence over time consists in nothing

but a temporal succession of your conscious mental states and the relevant relations between them. See, for instance, Bratman (2000; 2005), Parfit (1984), and Shoemaker (1984) for accounts along these lines. For criticism, see Olson (2007).

40 See Boghossian (2014, p. 5), Broome (2013), Burge (2013), and Wedgwood (2006, p. 669) for accounts of conscious reasoning that apply to you as a persisting conscious agent, rather than to proper parts of or systems within you. Note that you need not be aware of yourself *as* consciously reasoning.

41 Moreover, you also possess bodily capacities used during intentional bodily actions, so that you are the same conscious thinking agent who performs mental and bodily actions as one persisting physical object. Providing a unified explanation of bodily and mental action is a further virtue of the alternative account presented here, though one that must be explored on another occasion.

42 Thanks to the anonymous referees who reviewed earlier versions of this chapter on behalf of journals to which it was submitted. They provided valuable feedback. For helpful discussion, thanks to audiences at the Agency Workshop at Warwick, organized by Tom McClelland; the Metaphysical Society of America Meeting in Atlanta; the Works in Progress Seminar at Rice University, organized by Gwen Bradford; and the New Directions in the Study of Mind Seminar at Cambridge, hosted by Tim Crane. For insightful conversations about the ideas presented in this chapter or feedback on prior drafts, thanks to Andrei Buckareff, Lucy Campbell, Matt Duncan, Alexander Greenberg, David Hunter, Yair Levy, Lisa Miracchi Titus, Katia Samoilova, Josh Shepherd, Matt Soteriou, Daniel Telech, and Markos Valaris. Special thanks to the contributors of this volume. I'm deeply grateful for their patience and persistence with this book. Extra special thanks to Lisa Miracchi Titus for her work as co-editor. Without Lisa's efforts, and the diligent work of John Roman, Maja Sidzińska, and Jacqueline Mae Wallis, this book would not have seen the light of day. Finally, thanks to the University of Denver for its support of the research and travel that contributed to the development of this chapter, and the conference from which this edited volume emerged.

## References

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

Arpaly, N., & Schroeder, T. (2012). Deliberation and acting for reasons. *Philosophical Review*, *121*(2), 209–239.

Arpaly, N., & Schroeder, T. (2014). *In praise of desire*. New York: Oxford University Press.

Baird, B., Mrazek, M. D., Phillips, D. T., & Schoolar, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology: General*, *143*, 1972–1979.

Bargh, J. A. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control in social cognition. In R. S. Wyer, & K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1–4). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479.

Bayne (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 169–186). Cambridge, MA: MIT Press.
Bird, A. (2007). *Nature's metaphysics*. New York: Oxford University Press.
Bishop, J. (1989). *Natural agency: An essay on the causal theory of action*. New York: Cambridge University Press.
Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, *18*, 227–247.
Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*, 1–18.
Bouton, C. E., Shaikhouni, A., Annetta, N. V., Bockbrader, M. A., Friedenberg, D. A., Nielson, D. M. … Rezai, A. R. (2016). Restoring cortical control of functional movement in a human with quadriplegia. *Nature*, *533*(7602), 247–250.
Boyle, M. (2011). Making up your mind' and the activity of reason. *Philosophers' Imprint*, *11*(17), 1–24.
Bradford, G. (2015). *Achievement*. New York: Oxford University Press.
Brand, M. (1984). *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.
Bratman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
Bratman, M. E. (2000). Reflection, planning, and temporally extended agency. *Philosophical Review*, *109*, 35–61.
Bratman, M. E. (2005). Planning agency, autonomous agency. In J. S. Taylor (Ed.), *Personal autonomy*. New York: Cambridge University Press.
Brent, M. (2017). Agent causation as a solution to the problem of action. *Canadian Journal of Philosophy*, *47*(5), 656–673.
Broome, J. (2013). *Rationality through reasoning*. Chichester: Wiley-Blackwell.
Buckareff, A. A. (2005). How (not) to think about mental action. *Philosophical Explorations*, *8*(1), 83–89.
Buckareff, A. A. (2007). Mental overpopulation and mental action: Protecting intentions from mental birth control. *Canadian Journal of Philosophy*, *37*(1), 49–65.
Burge, T. (2013). *Cognition through understanding*. New York: Oxford University Press.
Chisholm, R. (1964/2003). Human freedom and the self. In G. Watson (Ed.), *Free will*. New York: Oxford University Press.
Chrisman, M. (2016). Epistemic normativity and cognitive agency. *Noûs*, *52*(3), 508–529.
Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.
Davidson, D. (1971). Agency. In R. Binkley, R. Bronaugh, & A. Marras (Eds.), *Agent, action, and reason*. Toronto: University of Toronto Press.
Davidson, D. (1978). Intending. In Y. Yovel (Ed.), *Philosophy of history and action*. Dordrecht: D. Reidel.
Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.
Ellis, B. (2001). *Scientific essentialism*. New York: Cambridge University Press.
Enç, B. (2003). *How we act: Causes, reasons, and intentions*. New York: Oxford University Press.
Evans, J. S. B. T. (2010). *Thinking twice: Two minds in one brain*. New York: Oxford University Press.

Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly*, *15*(2), 157–162.

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2018). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, *23*(2), 107–131.

Goldman, A. (1970). *A theory of human action*. Englewood Cliffs, NJ: Prentice Hall.

Goldman, A. (1971). The individuation of action. *Journal of Philosophy*, *68*(21), 761–774.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*(4), 495–525.

Harman, G. (1999). *Reasoning, meaning, and mind*. Oxford: Clarendon Press.

Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity*. Oxford: Blackwell Publishing.

Heil, J. (2012). *The universe as we find it*. Oxford: Clarendon.

Hill, C. L. M., & Updegraff, J. A. (2012). Mindfulness and its relationship to emotion regulation. *Emotion*, *12*, 81–89.

Hieronymi, P. (2009). Believing at will. *Canadian Journal of Philosophy, Supplementary*, *35*, 149–187.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J. … Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, *485*(7398), 372–375.

Hoffman, J., & Rosencrantz, G. (1994). *Substance among other categories*. Cambridge, UK: Cambridge University Press.

Holton, R. (2009). *Willing, wanting, waiting*. New York: Oxford University Press.

Hornsby, J. (2008). A disjunctive conception of acting for reasons. In A. Haddock, & F. Macpherson (Eds.), *Disjunctivism: Perception, action, knowledge*. Oxford: Oxford University Press.

Hornsby, J. (2004). Agency and actions. In J. Hyman, & H. Steward (Eds.), *Agency and action*. Cambridge, UK: Cambridge University Press.

Hornsby, J. (2013). Basic activity. *The Aristotelian Society Supplementary Volume*, *87*(1), 1–18.

Hunter, D. (2003). Is thinking an action? *Phenomenology and the Cognitive Sciences*, *2*(2), 33–148.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: Describing inner experience. *TRENDS in Cognitive Sciences*, *5*, 400–403.

Hurlburt, R. T., & Heavey, C. L. (2004). To beep or not to beep: Obtaining accurate reports about awareness. *Journal of Consciousness Studies*, *11*, 113–128.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kane, R. (1996). *The significance of free will*. New York: Oxford University Press.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

Levy, Y. (2016). Action unified. *The Philosophical Quarterly*, *66*(262), 65–83.

Li, Y., Liu, F., Zhang, Q., Liu, X., & Wei, P. (2018). The effect of mindfulness training on proactive and reactive cognitive control. *Frontiers in Psychology*, *9*, 1002.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*, 529–566.
Libet, B. (1987). Are the mental experiences of will and self-control significant for the performance of a voluntary act. *Behavioral and Brain Sciences*, *10*, 783–786.
Libet, B. (1989). The timing of a subjective experience. *Behavioral and Brain Sciences*, *12*, 183–185.
Lowe, E. J. (2008). *Personal agency: The metaphysics of mind and action*. Oxford: Oxford University Press.
Mast, F. W., Bamert, L., & Newby, N. (2007). Mind over matter? Imagined body movements and their neuronal correlates. In F. Mast, & L. Jäncke (Eds.), *Spatial processing in navigation, imagery and perception*. Boston: Springer.
Martin, C. B. (2008). *The mind in nature*. Oxford: Clarendon.
Mayr, E. (2011). *Understanding human agency*. New York: Oxford University Press.
McHugh, C., & Way, J. (2018). What is good reasoning? *Philosophy and Phenomenological Research*, *96*, 153–174.
Mele, A. (1997). Agency and mental action. *Philosophical Perspectives*, *11*, 231–249.
Mele, A. (2009). Mental actions: A case study. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions*. New York: Oxford University Press.
Molnar, G. (2003). In S. Mumford (Ed.), *Powers: A study in metaphysics*.. New York: Oxford University Press.
Mumford, S., & Anjum, R. L. (2011). *Getting causes from powers*. New York: Oxford University Press.
Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *8*, 435–450.
Nigel, T. (2017). Mental imagery. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). Retrieved from https://plato.stanford.edu/archives/spr2017/entries/mental-imagery/
O'Brien, L., & Soteriou, M. (Eds.). (2009). *Mental actions*. New York: Oxford University Press.
Olson, E. (2007). *What are we?* New York: Oxford University Press.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Paul, S. (2015). Doxastic self-control. *American Philosophical Quarterly*, *52*(2), 145–158.
Peacocke, C. (2007). Mental action and self-awareness (i. In B. McLaughlin, & J. D. Cohen (Eds.), *Contemporary debates in philosophy of mind*. Oxford: Blackwell.
Proust, J. (2001). A plea for mental acts. *Synthese*, *129*(1), 105–128.
Proust, J. (2013). *Philosophy of metacognition: Mental agency and self-awareness*. Oxford: Oxford University Press.
Ruben, D.-H. (1995). Mental overpopulation and the problem of action. *Journal of Philosophical Research*, *20*, 511–524.
Ruben, D.-H. (2008). Disjunctive theories of perception and action. In A. Haddock, & F. Macpherson (Eds.), *Disjunctivism: Perception, action, knowledge*. Oxford: Oxford University Press.
Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.

Schwitzgebel, E. (2016). Introspection. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Retrieved from https://plato.stanford.edu/archives/win2016/entries/introspection/

Sehon, S. R. (2005). *Teleological realism: Mind, agency, and explanation*. Cambridge, MA: MIT Press.

Sehon, S. R. (2016). *Free will and action explanation*. New York: Oxford University Press.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*(1), 99–124.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Shepherd, J. (2015). Deciding as intentional action: Control over decisions. *Australasian Journal of Philosophy*, *93*(2), 335–351.

Shepherd, J. (2016). Conscious action/zombie action. *Noûs*, *50*(2), 419–444.

Shoemaker, S. (1984). Personal identity: A materialist's account. In S. Shoemaker, & R. Swinburne (Eds.), *Personal identity*. Oxford: Blackwell.

Shoemaker, S. (1988). On knowing one's own mind. *Philosophical Perspectives*, *2*, 183–209.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *The Annual Review of Psychology*, *66*, 487–518.

Soteriou, M. (2013). *The mind's construction: The ontology of mind and mental action*. Oxford: Oxford University Press.

Strawson, G. (2003). Mental ballistics or the involuntariness of spontaneity. *Proceedings of the Aristotelian Society*, *103*, 227–256.

Upton, C., & Brent, M. (2019). Meditation and the scope of mental action. *Philosophical Psychology*, *32*(1), 52–71.

Valaris, M. (2016). What reasoning might be? *Synthese*, *194*, 2007–2024.

Velleman, J. D. (1992). What happens when someone acts. *Mind*, *101*(403), 461–481.

Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford: Oxford University Press.

Wedgwood, R. (2006). The normative force of reasoning. *Noûs*, *40*(4), 660–686.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*(7), 480–492.

Whittle, A. (2016). A defence of substance causation. *Journal of the American Philosophical Association*, *2*(1), 1–20.

Williams, N. E. (2014). Powers: Necessity and neighborhoods. *American Philosophical Quarterly*, *51*(4), 357–371.

Williams, N. E. (2017). Powerful perdurance: Linking powers with parts. In J. Jacobs (Ed.), *Causal powers*. New York: Oxford University Press.

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the will*. New York: Oxford University Press.

# 6 Inference as a Mental Act

*David Hunter*

> *Belief is the most difficult topic because it is so difficult to hold in view and correctly combine the psychological and the logical aspects.*
> (Anscombe, 1995, p. 26)

I will argue that a person is causally responsible for believing what she does. Through inference, she can sustain and change her perspective on the world. When she draws an inference, she causes herself to keep or to change her take on things. In a literal sense, she makes up her own mind as to how things are. And, I will suggest, she can do this voluntarily. It is in part because she is causally responsible for believing what she does that there are things that she ought to believe, and that what she believes can be to her credit or discredit. I won't pursue these ethical matters here, but will focus instead on the metaphysics that underpin them.[1]

This view of inference is quite natural, but it is obscured by familiar philosophical ideas about action, causation, and about inference itself. Full treatments of these ideas are beyond the scope of this chapter. So my modest aim is to describe in some detail a conception of inference that allows us to take literally the idea that a person can be the sustaining and originating cause of her own beliefs. The core of my view consists of three ideas about inference.

1 An act of inference is a causing, and not a cause, of believing.
2 In drawing an inference, the believer is the cause.
3 The believer does not cause the act of inference.

Here are two examples. Jones looks out the window and sees that it is raining hard.[2] She knows that a strong rain tends to melt snow quickly. So she concludes that the snow on the ski slopes will melt soon. Sarah believes that her son is not selling drugs, even though she has known for some time that the police are suspicious. But she finds the evidence to be inconclusive. She then meets with two police officers who present her with new photos and a signed witness statement. After carefully

DOI: 10.4324/9780429022579-7

considering this new evidence, Sarah continues to believe that her son is not selling drugs. The view I want to explore is that in cases like this, where a person forms or retains a belief after reflecting on evidence that is strong but not conclusive, she voluntarily causes herself to start or to continue believing something.

In Section 6.1, I will sketch the ideas about action, causation, and agency that underlie my view. In Section 6.2, I will consider Ryle's reasons for thinking that an inference is not an action but is, rather, the onset of a state of belief. In Section 6.3, I will consider an objection to the idea that in inference a person acts on herself. In Section 6.4, I will explore the roles that choice and desire play in inference. I will suggest that inference is voluntary but not always willing.

## 6.1 Actions, causings, and agents

My view of inference is a version of agent causation. While a full treatment of this is beyond the scope of this chapter, it will be helpful to flesh out the view a bit. The idea that acts are causings, or at least that some acts are causings, was defended by Judith Jarvis Thomson and by Kent Bach in the 1970s, who cited earlier work by Chisholm and Von Wright. And it has been developed, more recently, by Maria Alvarez and John Hyman.[3] The idea is simple enough. The act of scratching a table is not an event that causes the resulting scratch. Rather, the act is the causing of the scratch. The cause is whatever did the causing. A nail, perhaps, or little Billy who wielded it.

I am not assuming that all acts are causings. Some acts are defined in terms of a characteristic result. To scratch is to produce a scratch, to melt is to make something to melt, to bend is to cause something to bend, to push is to make something move, and etc. But not all actions have a characteristic result. Walking is an action, but there is no characteristic result that walking is in every case the causing of. The same is so for certain speech acts. Asserting is an action, but it has no characteristic result. Thomson (1977) noted this fact about actions, though she put it in linguistic terms by saying that not all verbs of action are causal verbs. But she also held that every action is either a causing (that is, reportable with a causal verb) or is done by doing something that is a causing. Walking, for instance, is done by moving one's legs and those acts are causings. Likewise, asserting is done by moving one's mouth or one's hands in certain ways, and those acts are causings.

Drawing an inference is, in this respect, unlike asserting something. Asserting something is always done *by* doing something else. One cannot simply assert something, one has to do it in a certain way, by making certain sounds or hand gestures. But drawing an inference is not like that. Maybe this can be put by calling inference a basic act. John Heil has objected to this.

> I am unable to shake the conviction that it is a mistake to regard the adoption of beliefs as actions in any sense, whether basic or non-basic. This conviction is not founded solely on the observation tendered earlier that the forming of beliefs is not something that can be accomplished by a sheer act of will, but on the evident fact that our beliefs seem to come to us rather than issuing from us. Paying a debt is something I can set out to do; believing something is not.
>
> (Heil, 1983, p. 358)

Heil assumes that action is basic only if involves an act of will. I have not spoken of acts of will. But I deny that to be an action an inference would have to be caused or generated by an act of will. So far as I can see, an inference need not have any cause. The believer does not cause the inference; she is the one inferring. And it is anyway a mistake to think that action is essentially tied to the will. When a bit of acid dissolves some rubber, the acid acts on the rubber, even though the acid has no will. Heil offers as further evidence the contention that our beliefs "come to us rather than issuing from us", but this seems to me to beg the question. And while he is right that we cannot believe or infer intentionally, this does not show that inference is not an action. The acid does not intentionally dissolve the rubber either.[4]

Heil is right that a person cannot draw an inference in order to achieve some further end or purpose. Inferences, in this sense, cannot be intentional. I will just take this as a datum, but it would be nice to know why. What is it about an act of inference that explains why it cannot be intentional? One might think it is because the act results in a belief state and that one cannot be in any state on purpose. But I am not sure this is right. For one thing, it seems one can be in some states on purpose. Jones is a vegan in order to help fight climate change. Simon is unemployed to focus on his art. What is more, the result of an intentional action can be intentional. The scratch Billy made on the car with the nail was intentional. And even if the resulting state of affairs is not intentional, the causing of it could be. Sarah scared the raccoon intentionally, even if the raccoon's fear was not intentional. So the fact that acts of inference result in states of affairs does not seem to me to explain why inferences cannot be intentional.

Perhaps it has to do with the fact that inferring requires taking the results of one's inference to be right and to be supported by one's reasons. Here is how Alan White put the idea.

> One can wrongly infer, but not infer in the belief that one is wrong, since the position taken up must be one which the person has come to believe to be related in a certain way to a previous position. This is why no question can arise of inferring voluntarily or by decision or resolution or on request.
>
> (White, 1971, pp. 291–292)[5]

I disagree with White on whether an inference can be voluntary, though this may only be a terminological matter. But I agree that we cannot infer in the belief that the result of our inference is false. I also agree that inference requires believing that the result is supported by our reasons, or at least that it requires not believing that the result is not supported by them. But why should we think that inference could be intentional only if an inference *could* result in a belief one took to be false or to be unsupported by one's reasons? Jones thinks that her reasons do support believing that the snow will melt quickly. She could surely draw that inference in the hope that by drawing it she would please her mom. Why couldn't she then make the inference *in order* to please her mom?

One might think that an inference cannot be an act if it cannot be intentional. This would be so if the following principle were true.

> S's V-ing is an act only if S can V intentionally.

But this principle is implausible. The tree outside my house scrapes my roof, which is an act, but the tree is not able to do anything intentionally. What about the following?

> V-ing is an act only if it is possible to V intentionally.

Even though trees can't scratch anything intentionally, people can. So this principle survives the objection I just considered. And if it is true, and if it is also true that an inference cannot be intentional, then it would follow that inferences are not acts. But is this principle true? A pine tree can produce cones, and I take it that producing cones is an act. But must it then be possible to produce cones intentionally? (I can intentionally make a tree produce cones, but this is different.) I am not sure. And, anyway, what would explain such a principle? In effect, it holds that intentionality is essential to action. But on the view I have been exploring, it is causation and not intentionality that is fundamental to action. The idea is that to act is to cause or bring about some change. Some actions are intentional, just as some are voluntary and some are reluctant. But there is nothing in the idea of action itself that requires that any act be potentially intentional. And while intentional acts done by people are of interest to ethics, ethics is just as interested in what people do inadvertently and by mistake. Indeed, in considering whether a person ought to have done something, it often matters little whether the person did it intentionally or what her intentions were.[6] So I find it hard to see what considerations could support such a principle.

And notice just how extreme it is. Some of our capacities are ones that, while we cannot exercise them intentionally, we might have been able to. Most of us are able to weep, but few of us can weep intentionally. We can imagine that with training and practice we might become able

to weep intentionally, in order to get something we want. Maybe good actors can already do this. So it is plausible to think that *some* of our current powers are ones we could have been able to exercise intentionally even though at present we cannot. But the principle I am considering says that for *every* power we have, we could have been able to exercise it intentionally, and this just in virtue of its being a power. This strikes me as implausible.

Finally, it seems to me that if we are to take at face value the idea that in inference we are making up our own minds, then we need to reject that principle linking action and intentionality. In effect, insisting on the principle begs the question against my view.

## 6.2 Inference and the onset of believing

Gilbert Ryle argued that an inference is the onset of a state of affairs and not an action at all. Considering his reasons will help clarify my view. And, as we will see, Ryle was not himself completely convinced by them.

Here is how Ryle objected to the idea that inference is an action.

> We saw that there was some sort of incongruity in describing someone as being at a time and for a period engaged in passing from premises to a conclusion. 'Inferring' is not used to denote either a slowish or a quickish process. 'I began to deduce, but had not time to finish' is not the sort of thing that can significantly be said. In recognition of this sort of incongruity, some theorists like to describe inferring as an instantaneous operation, one which, like a glimpse or a flash, is completed as soon as it is begun. But this is the wrong sort of story. *The reason why we cannot describe drawing a conclusion as a slowish or quickish passage is not that it is a 'Hey, presto' passage, but that it is not a passage at all.*... [R]eaching a conclusion, like arriving in London, solving an anagram and checkmating the king, is not the sort of thing that can be described as gradual, quick or instantaneous.
>
> (Ryle, 1949, pp. 301–302; italics added)

> 'Conclude', 'deduce' and 'prove', like 'checkmate', 'score', 'invent' and 'arrive', are, in their *primary* uses what I have called 'got it' verbs, and while a person's publications, or other exploitations of what he has got, may take much or little time, his *transition* from not yet having got it to having now got it cannot be qualified by epithets of rapidity.
>
> (Ryle, 1949, p. 276; italics added)

Ryle's concern, in that first passage, is with the idea that an inference has a duration. He thinks this cannot be right, since if it were we should be able to sensibly ask how long one took, and whether it was slow or

quick. But, as he correctly notes, these questions make no sense. He then rejects the suggestion that an inference has an instantaneous duration, completed as soon as it is begun.

On my view, an inference is a causing and I can accept that inferences have no duration. For causings have no duration. We need to distinguish how long an exercise of a capacity *lasts* from how long the exercising *takes*. The drop of acetone melted the bit of rubber for two minutes before it was wiped off. The capacity was exercised for two minutes, but the exercising itself was neither fast nor slow, for it took no time. Likewise, a person might hold their standing position for 20 seconds. The holding lasted for 20 seconds, but the holding itself took no time. If the holding itself took time, we should be able to ask whether it was a quick holding or a slow one, and whether it got faster near the end. And one could be asked to speed up or slow down one's holding of a position. These questions make no sense, because holding one's position is causing it not to change, and causings don't take time.

After Ryle rejects the idea that an inference is an event, he settles on the idea that an inference is the onset of a state of affairs. The idea is that when Jones infers that the snow on the slopes will melt soon, that inferring is the onset or start of her believing that the snow on the slopes will melt soon. It is the onset of a certain mental state of affairs. I agree that onsets have no duration, and so it would make little sense to ask whether the onset was quick or slow. I also agree that onsets are not acts. They seem, rather, to be a possible result of an act.

Thomson agrees that an inference is the onset of a state of affairs. She groups inferring with deducing, discovering, and remembering.

> Alfred's recalling this and such is an onset of the state of affairs that is his remembering this and such; Bert's recognizing so and so is an onset of the state of affairs that is his being aware of who or what so and so is; Charles's noticing thus and such is an onset of the state of affairs that is his being aware of thus and such; etc.
>
> (Thomson, 1977, p. 230)

(She considers onsets to be events, and so does not share Ryle's discomfort with the idea of an instantaneous event.) I am inclined to agree that a sentence like

> Peter discovered his husband's infidelity.

can report the onset of Peter's knowing about the infidelity, and that a sentence like

> Simon remembered where he left his keys.

can report the onset of Simon's awareness of the key's location. But I don't think that inferring properly belongs in this group. For one thing, the connection to reasons is quite different. One draws an inference in light of certain reasons one has, but one does not discover or remember something in light of a reason one already has. What is more, one can try or attempt to remember or discover or find something, but one cannot try or attempt to infer something. Here is how Alan White put the point, though he spoke of achievements instead of onsets.

> Inferences are not achievements or arrivals, because, unlike discoveries (and also unlike deductions) they are not something we can try to, promise or resolve to make or can manage to obtain. We can not use means and methods or rely on luck to infer something. We can ask someone what he would infer from the evidence but not how he would infer. An examination paper could ask the candidates to solve, prove, find, discover or even deduce so and so, but it could not sensibly ask them to infer.
>
> (White, 1971, p. 291)[7]

So I am not convinced that an inference is the onset of a state of affairs.

I am also not sure that Ryle was really convinced either. In one place, he suggests that an inference is a *performance* (Ryle, 1946, p. 22). In another, he calls it an *operation* (Ryle, 1949, p. 274), and in a third, he suggests it is the *result* of the performance or operation (Ryle, 1949, p. 260). Though Ryle's views on the nature of inference are, at the end of the day, a bit unclear, we know what sort of view he was opposing. He was opposing what he considered the 'para-mechanical idea' of inference as a mental act or process that causes a mental state.

> Finding premises and conclusions among the elements of published theories, [the Epistemologists] postulate separate, antecedent, 'cognitive acts' of judging, and finding arguments among the elements of published theories, they postulate antecedent *processes of moving to the 'cognising' of conclusions from the 'cognising' of premises*. I hope to show that these separate intellectual processes postulated by epistemologists are para-mechanical dramatisations of the classified elements of achieved and expounded theories.
>
> (Ryle, 1949, p. 291; italics added)

Ryle is opposed to the idea that inferring is a process that is separate from the resulting belief. A lot rides, of course, on what he meant by 'separate.' But a natural interpretation is that on the para-mechanical view inferring is an event or process that causes the resulting belief.

Some recent accounts of inference strike me as versions of a para-mechanical view. Robert Audi says that an inference "produces" a belief

as a "process of passing from one or more premises to a conclusion" (Audi, 1986, p. 31). According to John Broome, "[r]easoning is a mental process through which some attitudes of yours…give rise to a new attitude of yours" (Broome, 2014, p. 622). Paul Boghossian says that "[i]t's not sufficient for my judging (1) and (2) to cause me to judge (3) for this to be inference. The premises judgments need to have caused the conclusion judgments 'in the right way'" (Boghossian, 2014, p. 3). I agree with Ryle in opposing such para-mechanical views.

## 6.3 Acting on oneself

On my view, in inference, a believer causes herself to believe something, and so acts on herself. This might seem to be the exact opposite of something Stuart Hampshire says.

> The man who changes his mind, in response to evidence of the truth of a proposition, does not act upon himself; nor does he bring about an effect.
>
> (Hampshire, 1975, p. 100)

Hampshire says that when a person changes her mind in response to evidence, she does not act on herself. When Jones concluded that the snow on the slopes will melt soon, she did not, Hampshire seems to be saying, *make up her mind*, at least not in a straightforward causal sense of that phrase. But I think Hampshire's view may be more nuanced than this. We need to distinguish two ways of acting on oneself. And we need to avoid the temptation to hypostasize believing.

One way to act on oneself is to do something that causes or brings about a change in oneself. So, for instance, Jones cuts her nails with a pair of clippers. In doing that, she acted on her nails (and so on herself) by closing the clippers on them. Here she is the agent and patient of the action. She is an agent since she did the cutting. She is patient since she was cut. A person can cause herself to believe something in this sort of way. Jones can make herself believe that the glass of beer is empty by emptying it. Here she is again both agent and patient. She is an agent because she did something that caused her to believe the glass is empty. And she is patient since her beliefs were changed. But, I take it, this is not a case of making up one's mind through reasoning.

Another way to act on oneself is to do something that is *itself* a changing of oneself. So, for instance, Jones can change her bodily position by crossing her legs. Here again, she is both agent and patient. She is the agent since she crossed her legs. She is patient since her legs were crossed. But crossing one's legs is not an action that causes or brings about a change in one's position. Crossing one's legs *is* changing one's position.

On my view, making an inference is acting on oneself in this second way. It is not an action that has *as a consequence* that one believes something. It is not acting on oneself in the way that cutting one's nails is acting on oneself. Rather, it is acting on oneself in the way that crossing one's legs is acting on oneself. Making an inference is changing or sustaining one's state of believing. Perhaps this is what Hampshire had in mind in the passage I quoted.

The tendency to hypostasize believing, to treat belief states as particulars in the same ontological category as fingers and nails, can make this view difficult to see. For that tendency encourages the idea that if inference is an action it must be like clipping one's nails. Matthew Chrisman seems to have this sort of view in mind.

> What is involved in maintaining a system of beliefs? As the verb phrase suggests, it is dynamic rather than static. Maintaining something (e.g., a flowerbed) can be a reasonable answer to the question "What are you doing?" … More specifically, as I am thinking of it, maintaining a system of beliefs involves *examining* and *adjusting* existing *beliefs* in light of newly acquired beliefs or propositions assumed to be true for various purposes (e.g., by raising or lowering one's credence in the old beliefs, or by *reconceiving* the inferential/evidential relations between beliefs if they seem to be in tension under various suppositions). It can also involve seeking out new beliefs — e.g., by investigation or deliberation — when one's system of beliefs leaves some important question open or some strongly held belief apparently unsupported by other beliefs.
>
> (Chrisman, 2016, p. 16; italics added)

Three ideas about believing that I have critiqued elsewhere are at work in this passage.[8] First, states of believing are treated as particulars in the same category as plants (hence the plural 'beliefs'); they are taken to have semantic properties (for they are taken to bear inferential relations to one another), and they are considered to be things a person can, though reasoning, act on and adjust. That is the point of Chrisman's analogy that maintaining one's beliefs is like maintaining the geraniums in one's flowerbed. Pictured this way, acting on one's beliefs would be like acting on one's nails.[9]

Chrisman is responding to Matt Boyle, who has argued that we will have trouble understanding a person's responsibility – their agential control – for believing what they do if we think of believing as a state.[10] As Boyle sees it, on the state view,

> [i]f we exercise agential control over our beliefs, this must consist in our performing occurrent acts of judgment which give rise to new beliefs, or cause extant beliefs to be modified. Beliefs can at most

> "store" the results of such acts. So a person's agency can get no nearer to her beliefs than to touch them at their edges, so to speak.
>
> (Boyle, 2009, p. 121)

Boyle's thought is that while the state view of belief can allow acts of judging or affirming, their role must be limited to being acts that cause or change states of belief. But this, Boyle charges, distorts our relation to our beliefs.

> [The state view] appears to leave us responsible only for *looking* after our beliefs, in something like the way I may be responsible for looking after my bicycle. I have chosen to acquire this bicycle, and I can take steps to ensure that it is in good condition, that it is not left in a bad spot, etc. I am responsible for it as something I can assess and act upon, something in my care. I am not responsible for it, however, in the way, I am responsible for my own intentional actions. My actions stand in a more intimate relation to me: they are not things I control by acting *on* them; they are my doings themselves.
>
> (Boyle, 2009, p. 121; italics in original)

Of course, no one thinks that beliefs are like files on hard drives or recordings on machines, or that our relation to them is like our relation to our bicycles. Everybody should agree that believing is not like that. The purpose of Boyle's caricature is to force us to say how believing is different.

I think we can make a start on seeing the difference if we consider an analogy between believing and owning. Jones owns many books, and she can reflect on and organize what she owns. But she does this, not by reflecting on and organizing her *possessings*, but by reflecting on and organizing her *possessions*. Her possessings are not things at all, let alone things she can causally interact with. Something similar is true for reflection on how one takes things to be. When Jones considers how things are, her attention is directed at what she believes not at her believings. She is attending to the possibilities she thinks obtain and to those that remain open. She is not attending to her mental states.

There is of course an important disanalogy between believing and owning. The objects of ownership can be physical things that one can causally interact with. Jones can move and pile up her books. But the objects of belief are not things one can causally interact with. One cannot move or arrange them. In this respect, a better analogy is with bodily position. On my view, to believe something is to be in a certain position with respect to how things are and might have been, with respect to a range of possibilities. We can compare this to occupying a position in space. Jones can maintain or adjust her position by moving her legs and

arms. This is not a matter of doing something that causes a change in the positions of her arms and legs. It is, rather, doing something that consists in changing her position. This is how we should think of reasoning. A person maintains or changes her position on the way things are, not by doing something that sustains or changes that position, but by doing something that is itself a sustaining or changing of it. She is the agent of the action, and also its patient, and is causally responsible for the result.[11]

## 6.4 Inference, choice, and desire

I said that an inference cannot be intentional. But it is a separate matter whether an inference can be voluntary. I am inclined to think it can, though getting clear on the relevant sense of 'voluntary' is not easy.[12]

A person does something voluntarily if she chooses to do it.[13] This is just a sufficient condition, and it applies only to what a person does, not to what she may voluntarily undergo or to ways she may voluntarily be. And it leaves open what role the choosing plays. If choosing to do something requires that one choose to do it before one does it, then inference is not voluntary in that way, for a person cannot choose to draw an inference. We cannot decide or make up our minds in advance to believe something.

But the idea of choice also suggests that the person had alternatives. A person who does something by choice was not coerced into doing it. Coercion is itself a nuanced matter, for a person can be coerced into doing something, and so not do it voluntarily, even when she could have refrained from doing it. A person who does something in response to a realistic threat of serious violence does not do it voluntarily, even if she could have refrained. Still, if a person does something, and could have reasonably refrained from doing it, and did not act from ignorance, then she did the thing voluntarily.[14]

Could an inference be voluntary in this second way? Some have said that when a person draws an inference her drawing it is compelled or determined by the evidence she has. This suggests that the person has no alternative but to draw the inference and so could not refrain. If so, then an inference would not be voluntary in this second way. Here is how David Owens puts the point. His example involves John's being the murderer.

> [W]hat directly determines how we think about John are his bloody shirt and absence from work. I do decide to attend to these things, but once that decision is made, the evidence takes over and I lose control. It might be argued that I held off from forming a view on the basis of his pleasant demeanor alone and waited for more evidence. Doesn't this amount to an exercise of control over what I

> believe? But all that occurred was that his agreeable countenance proved insufficient to close my mind on the matter, to eliminate doubt about his innocence, and so I set off once more in search of evidence. In the end, it is the *world* which determines what (and whether) I believe, not me.
>
> (Owens, 2000, p. 12)[15]

According to Owens, when a person draws an inference in light of the evidence she has, her evidence leaves her no alternative in the matter. John Heil agrees, saying that in inference believers are "largely at the mercy of their belief-forming equipment" (Heil, 1983, p. 357). The idea that inference involves the operation of autonomous belief-forming equipment is reminiscent of the para-mechanical account opposed by Ryle.

Importantly, even if Owens and Heil are right, this would not show that an inference is not an action done by the believer. Think again about the drop of acetone. When it touches the rubber its power to dissolve the rubber is activated, and there is no alternative. (Barring masks, which we can set aside.) The conditions that are necessary for the acetone's exercise of its capacity to melt rubber are also sufficient for that exercise. Still, the acetone melted the rubber, and its melting of the rubber was an exercise of its capacity to melt it. If Owens is right, a person's having adequate evidence is sufficient for the exercise of her capacity to infer. But this would not entail that the person did not herself make the inference. So Owen's view is compatible with my idea that an inference is a causing of a believing by a person.

Still, it seems to me that Owens overstates the power of a person's evidence. We make inferences when the evidence we have is adequate but not conclusive. In such cases, it seems to me, we are free both to draw the inference and to refrain from drawing it. Drawing it would be reasonable, because the evidence we have is adequate. But refraining from drawing it would also be reasonable, because that evidence is not conclusive. Whether we draw it is up to us and is thus voluntary in the second way I specified above.

Trudy Govier defends this view.[16] She notes that discussions of inference tend to focus on cases where the evidence strongly favours one thing.

> The restriction of decision and choice to this kind of context seems to me to be mistaken. I think that this mistake is a result of concentrating too much on cases where there is no problem about what to believe; – where evidence is taken in and is so straightforward in its import that there is no need for conscious reflection. And of course, many (perhaps most) cases in which people come to believe things are like this. But not all are, and this is important. When one

> has insufficient or ambiguous evidence, or when one has to decide whether to go and seek evidence, and if so, what kind, there is a conscious reflection concerning what to believe.
>
> (Govier, 1976, pp. 653–654)

There are two elements here. One is that it would be reasonable to believe the thing in light of that evidence, but also reasonable to continue to suspend belief. (Since suspending belief is simply maintaining one's view, we should think of suspending belief as *sustaining* belief.) The second element is that the belief is formed in response to that evidence, and not in response to something the person wants. We should not think of voluntary inference as requiring that the believing be based on non-epistemic considerations. We should allow that the result of a voluntary inference could be perfectly reasonable.[17]

The examples I gave in Section 6.1 involve these same two elements. Jones has good reason to think the snow will melt quickly, but her evidence is not conclusive. It would be reasonable for her to believe it, but also reasonable for her to suspend judgment. In light of that evidence, she infers that it will melt soon. Making that inference is taking the evidence to be adequate. Sarah has good evidence that her son is dealing drugs, but it is not conclusive. It would be reasonable for her to believe it, but also reasonable for her to sustain her current beliefs. In light of that evidence, she sustains her view rather than changing it. Not concluding that he is dealing drugs is taking the evidence to be not conclusive.[18]

It seems to me that cases of this sort are commonplace.[19] We often have good but not conclusive evidence that some possibility obtains. Believing that it does would be reasonable. Suspending belief would also be reasonable. In such cases, it is up to the believer whether she changes or sustains her belief state. Drawing the inference is taking the evidence to be adequate. Suspending belief is taking it to be inconclusive. Both options are rationally open to the believer, who is free to change or sustain her take on the world. The result, in either case, will be due to the believer's exercise of her inferential power.[20]

This point is easily obscured if we model inference on formal deduction. Some cases of deduction make voluntary inference seem suspiciously easy while others make it seem invariably irrational. Disjunction introduction, where the disjuncts are logically independent, is a case of the first. Suppose Stephanie says the following.

> It is snowing; so either it is snowing or Toronto is in Canada; so either it is snowing or Toronto is in Canada or the 504 streetcar will be late, and so on.

If one thinks of this as a case of inference, then one will be tempted to think that it is entirely up to the believer how long to continue drawing

the inferences. Stephanie might decide to keep going, drawing more and more consequences from her initial premise. And she might continue until she chooses to stop. She might even do so intentionally, in order to bother her brother, or to win a bet, or just because she finds it amusing. Thinking of this as a case of inference will make it seem all too easy to infer voluntarily. But this is not a case of inference at all. At least, it is not what I take inference to be. I take it that an inference requires changing or sustaining one's view of which possibilities obtain. But in our story, Stephanie is not changing or sustaining her mind about which possibilities obtain. She is not adding to her map of the world. Nor is she sustaining it. She is simply formulating new ways to state what she already believes.

This case of disjunction introduction makes rational voluntary inference look suspiciously easy. Cases of modus ponens can make it look invariably irrational. Suppose Margaret believes that if it is raining, then the snow will melt. And suppose she starts believing that it is raining. She might then report her reasoning by telling us this.

> If it is raining the snow will melt, and it *is* raining; so, the snow will melt.[21]

One might think that if she performed an inference here it would have to be located in between believing the premises and believing the conclusion in her report. That is, one might think that Margaret inferred that the snow will melt *after* already believing the conditional and its antecedent. But it is hard to see how such an inference could have been rationally voluntary. Once she starts believing the second premise she has no rational option but to believe the conclusion. If it were up to her whether to draw it, then this would be a freedom to be irrational.

But this description of the case is misleading, for it locates the inference in the wrong place. I agree that Margaret must believe the conclusion, given that she believes the premises. But this is because *in believing* those premises she already believes the conclusion. She did not form a further belief in addition to the conditional and its antecedent. Given that she believed the conditional, in coming to believe its antecedent she came to believe its consequent. It is misleading to suggest that an additional step is needed in her reasoning, to get from belief in the premises to belief in the conclusion. Think of this in terms of adding information to a novel. Suppose the author has already stated that in the novel's world snow melts whenever it rains. If she then adds to the story that it is raining, she therein adds that snow is melting. She does not need to add a further sentence saying that the snow is melting. That is already so in the world of the novel.

Margaret may well have drawn an inference in this story. Whether she did depends on how she came to believe that it is raining. Suppose

she was listening to the radio while preparing her dinner and heard the announcer say that it is raining. This is some reason to think it is raining, but it is not conclusive. Margaret trusts the weather reporter but also knows the report is sometimes about a distant city. This is exactly like the case Govier describes. Neither belief nor suspension of belief is rationally mandatory. Suppose Margaret concludes that it is raining. On my view, this was a voluntary inference. And, on my view, she therein also inferred that the snow will melt. And this inference too was voluntary. Indeed, it was the same inference. So the case of modus ponens can make voluntary inference seem impossible, but only by mis-locating the inference.

Philip Nickel agrees that inference can be voluntary. But in discussing its scope, he too seems to mis-locate inference. He says the following.

> My main aim is to establish the intuitive plausibility of the view that there are some instances of doxastic willing, and defend the view against prevalent objections. But for all I have said here, we are often not in voluntary control of our beliefs, for in many cases there may be only one reasonable option. When I see a dog race toward me, I do not feel free to believe that there is no dog (or no animal) racing toward me, nor do I feel free to suspend judgment. It seems I come to believe it regardless of my doxastic character traits, because there is, at the end of the day, only one doxastic option.
>
> (Nickel, 2010, p. 331)

As he describes the case, there is a gap between his seeing the dog race towards him and his believing that the dog is racing towards him, and during this gap, he forms the belief that it is. But I find this hard to understand. Normally, if a person *sees* that a dog is racing towards her then she *knows* that it is. Seeing is a way of knowing. And if she knows it, then she believes it, since knowing requires believing. Being presented with a fact, with a way that things are, just is knowing them to be that way. No inference is needed, because none is possible. So there is no room for a gap between seeing that the dog is racing towards him and believing that it is. It seems to me that Nickel is wrong to think that this is a case of *non-voluntary* inference, for it is not a case of inference at all.

We can easily adjust the case to make it one of inference. Suppose that Philip does not see that the dog is racing towards him. He sees the dog running but is uncertain about its path. It might be aiming to attack him or it might be aiming to attack something to the side. Given this, it would be reasonable for Philip to believe that the dog is running towards him, but also reasonable for him to suspend belief. Because he does not see *that* the dog is running towards him, he does not yet believe that it is. It is up to him whether to believe it or whether to continue to suspend belief. So when we adjust the case to make it involve an inference, we see that it is, in the sense at issue, voluntary after all.

I have been suggesting that an inference can be voluntary in a sense I have tried to specify. One might think that an act can be voluntary only if it can be intentional. If so, then either I am wrong to think that inference can be voluntary, or else I am wrong to think that it cannot be intentional. But I question this link between the voluntary and the intentional. No doubt many acts that can be voluntary can also be intentional. But why should every act be like that? As Hyman (2015) argues, the concepts of the voluntary and the intentional are keyed to very different aspects of our mental lives. Whether an act is voluntary depends on whether a person was coerced or acted from ignorance. Whether it is intentional depends on whether she did it in order to satisfy a desire she has. Why couldn't there be an act that a person could voluntarily do, in light of certain reasons, but not ever do it in order to satisfy a desire? It seems to me that inference would be precisely such an act.

I have argued that when a person draws an inference, her action is, in a sense, I tried to specify, voluntary. But it does not follow that she wanted to make the inference. A person can do something voluntarily while wishing that she were not doing it, and while preferring that she not be doing it. In this sense, her doing it can be unwilling. Sarah might wish she had some better alternative than to report her son to the authorities. When she finally does, she does so intentionally and voluntarily but also, in an important sense, unwillingly. The same psychic complexity is possible in the theoretical realm. A person can draw an inference reluctantly and even unwillingly. There are two sorts of cases.

In one, the person is reluctant because she does not want what she has concluded to be the case. My example of Jones is like that. She spent a considerable amount on the ski trip and made many personal sacrifices to get there. She will be very disappointed if she is not able to ski. Admitting it would force her to modify her plans, something she did not want to face. And so, she unhappily came to the conclusion that the snow on the slopes will melt soon. By contrast, others in her group were much less committed to the skiing and were looking forward to spending time with friends. They were much more willing to conclude that the snow on the slopes would melt soon.

A second sort of case is more interesting. In it, the person is reluctant to draw the conclusion because she wishes she were more able to resist drawing it. A slight alteration of my Sarah case is like this. She feels that a truly loving mother would always give her children the benefit of the doubt and would not be moved by the sort of evidence the police might provide. As she surveys the new evidence, she knows that accepting that her son is selling drugs will make her feel ashamed, and she dreads the look of betrayal in her son's eyes when he confronts her. Still, when she sees the evidence the police provide, she concedes that he is in fact selling drugs. But her concession is reluctant and unwilling. Of course, she would also much prefer that

he not be a drug dealer. But what is distinctive in her case is that she also wishes that she were the sort of mother who could continue to believe in her son's innocence even in the face of all that evidence. That's the sort of loving mother she wishes she were. The prospect of failing in her love for her son makes her reluctant to conclude that he is a drug dealer.[22]

## Notes

1. The work in this chapter develops ideas in (Hunter, 2018b) and is part of a larger project currently in preparation.
2. This example is slightly adapted from John Broome (2014), who adapts it from Boghossian (2014). John Broome agrees that an inference is something a person does. "Some processes of reasoning are 'active', by which I mean they are done by the reasoner. Active reasoning is something we do, like *eating*, rather than something that just happens in us, like *digesting*" (Broome, 2014, p. 622; italics added). But I will argue that unlike eating, inferring is not done by doing something else and cannot be intentional.
3. See Thomson (1977, ch. XVIII), Bach (1980), and Alvarez & Hyman (1998).
4. The idea that acts are causings is compatible with different views about what sorts of entities can be agents and about the role of events in causation. Thomson (1977) says that events, in addition to individuals and bits of substances, can be causes, whereas Hyman (2015) denies that events can be agents. I won't pursue this here.
5. Notice that these considerations also explain why inference cannot, as he sees it, be voluntary. I turn to this below.
6. This is a point that Thomson makes (1977, p. 253).
7. White says that "[t]o infer is neither to journey towards, nor to arrive at or to be in a certain position; it is to take up, to accept or to *change to* a position" (White, 1971, p. 291; italics added). This seems compatible with the idea that an inference is the causing of the change. For more discussion, see Rumfitt (2012).
8. What I say here is deeply indebted to the important work in Marcus (2006, 2009). I develop the idea that beliefs are states a person is in and not states inside a person in Hunter (2001, 2018b). Similar views are in Kenny (1989) and Steward (1997).
9. This idea also informs standard accounts of deliberation. According to Nishi Shah, reasoning involves acting on and attending to one's mental states: "normally when we *reflect on our attitudes*, we do not merely come to know what we in fact believe or intend; we determine what we shall believe or intend" (Shah, 2013, p. 311; italics added). John Broome resists the idea that reasoning involves attending to our mental states. He holds onto the idea that it involves attending to objects: in reasoning "you *operate on* the contents of your premise-belief, following a rule, to construct a conclusion, which is the content of a new belief of yours that you acquire in the process" (Broome, 2014, p. 624; italics added).
10. Pamela Hieronymi shares Boyle's concerns. See, for instance, Hieronymi (2009).
11. I develop this view of believing in a book manuscript, in preparation.
12. Boghossian says it is voluntary (2014, p. 3), though he does not elaborate.
13. What follows relies on the discussion of voluntariness in Hyman (2015).

14. The knowledge condition in general is complex. Does it require knowing or can believing be sufficient? Does the absence of knowledge or belief make the action involuntary or just non-voluntary? In the case of inference, anyway, a person must know both that she is drawing the inference and something about the reasons she has for drawing it. This follows, I think, from the fact that a person who believes something knows that she does. And this knowledge condition is interestingly similar to the knowledge condition in the case of intentional action. I discuss all of this in work currently in preparation.
15. My thoughts on Heil and Owens and on doxastic agency are influenced by Gary Watson's (2003).
16. So did Roderick Chisholm (1968, p. 224). For an excellent discussion of the history of voluntarism about belief, see Pojman (1986). More recently, the idea that an inference is voluntary is defended by Philip Nickel (2010). I discuss his views below.
17. Chisholm (1968) holds that if it is reasonable for a person to believe something given her evidence, then her believing it is either morally permitted or morally required. Miriam McCormick (2015), among others, rejects this evidentialist view and argues that a person may be morally permitted or even required to believe something that conflicts with her evidence. In Hunter (2018a) I argue for a middle position: what it is reasonable for a person to believe always depends on her evidence, but what she ought to believe always depends on what she ought to know and this, in turn, depends on what she ought to do, feel, think and on how she ought to be.
18. But doesn't a person's epistemic character play a role? Isn't her inference determined by her evidence together with her character? I don't think so. People do have epistemic character traits, such as being hasty or cautious. But these traits are partly the products of repeated acts of inference, and do not cause them. A person's selfishness does not make them act selfishly.
19. William Alston, who famously argued against a deontological conception of epistemic justification on the grounds that belief is not voluntary, nonetheless allowed that in cases like Sarah's belief might be voluntary (Alston, 1988, p. 265). But, he insisted, such cases are so rare that no plausible conception of epistemic justification should be rested on them. I disagree about their rarity.
20. Roger White (2005) sees an incoherence in this idea. But, so far as I can tell, it depends on thinking of voluntary inference as involving making a choice, which I deny. For discussion of this, see Nickel (forthcoming).
21. The example is from Broome (2014). My variation on it is not a case of modus ponens. For another paper that takes deductive reasoning as the model of inference, see Hlobil (2019).
22. For more on this sort of case, see my Hunter (2011).

## References

Alvarez, M., & Hyman, J. (1998). Agents and their acts. *Philosophy*, *73*, 219–245.

Alston, W. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives*, *2*, 257–299.

Anscombe, G. E. M. (1995). Practical inference. In R. Hursthouse, G. Lawrence, & W. Quinn (Eds.), *Virtues and reasons* (pp. 1–34). Oxford: Clarendon Press.

Audi, R. (1986). Belief, reason, and inference. *Philosophical Topics*, *14*(1), 27–65.

Bach, K. (1980). Actions are not events. *Mind*, *89*(353), 114–120.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*, 1–18.
Boyle, M. (2009). Active belief. *Canadian Journal of Philosophy, Supplementary Volume*, *35*, 119–147.
Broome, J. (2014). Normativity in reasoning. *Pacific Philosophical Quarterly*, *95*, 622–633.
Chisholm, R. (1968). Lewis' ethics of belief. In P. A. Schlipp (Ed.), *The philosophy of c.I. Lewis* (pp. 223–242). La Salle, IL: Open Court.
Chrisman, M. (2016). Epistemic normativity and cognitive agency. *Noûs*, *52*(3), 508–529.
Govier, T. (1976). Belief, values, and the will. *Dialogue*, *15*, 642–663.
Hampshire, S. (1975). *Freedom of the individual*. London: Chatto and Windus.
Heil, J. (1983). Doxastic agency. *Philosophical Studies*, *43*(3), 355–364.
Hieronymi, P. (2009). Believing at will. *Canadian Journal of Philosophy, Supplementary Volume*, *35*, 135–187.
Hlobil, U. (2019). Inferring by attaching force. *Australasian Journal of Philosophy*, *97*(4), 701–714.
Hunter, D. (2001). Mind-brain identity and the nature of states. *Australasian Journal of Philosophy*, *79*(3), 366–376.
Hunter, D. (2011). Alienated belief. *Dialectica*, *65*(2), 221–240.
Hunter, D. (2018a). Directives for knowledge and belief. In C. McHugh, J. Way, & D. Whiting (Eds.), *Normativity: Epistemic and practical* (pp. 68–89). Oxford: Oxford University Press.
Hunter, D. (2018b). The metaphysics of responsible believing. *Manuscrito*, *41*(4), 255–285.
Hyman, J. (2015). *Action, knowledge, and will*. Oxford: Oxford University Press.
Kenny, A. (1989). *The metaphysics of mind*. Oxford: Oxford University Press.
Marcus, E. (2006). Events, sortals, and the mind-body problem. *Synthese*, *150*, 99–129.
Marcus, E. (2009). Why there are no token states. *Journal of Philosophical Research*, *34*, 215–241.
McCormick, M. (2015). *Believing against the evidence*. New York: Routledge.
Nickel, P. (2010). Voluntary belief on a reasonable basis. *Philosophy and Phenomenological Research*, *81*(2), 312–334.
Nickel, P. (forthcoming). Freedom through skepticism. English version of 'Vrijheid door sceptisicism', in Algemeen Nederlands Tijdschrift voor Wijsbegeerte.
Owens, D. (2000). *Reason without freedom*. London: Routledge.
Pojman, L. (1986). *Religious belief and the will*. London: Routledge, Kegan, and Paul.
Rumfitt, I. (2012). Inference, deduction, logic. In J. Bengson, & M. A. Moffett (Eds.), *Knowing how: Essays on knowledge, mind, and action* (pp. 334–360). Oxford: Oxford University Press.
Ryle, G. (1946). Why are the calculuses of logic and arithmetic applicable to reality? *Proceedings of the Aristotelian Society*, *20*, 20–60.
Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.
Shah, N. (2013). Why we reason the way we do. *Philosophical Issues*, *23*, 311–325.
Steward, H. (1997). *The ontology of mind*. Oxford: Oxford University Press.
Thomson, J. (1977). *Acts and other events*. Ithaca, NY: Cornell University Press.

Watson, G. (2003). The work of the will. In S. Stroud, & C. Tappolet (Eds.), *Weakness of will and practical irrationality* (pp. 172–200). New York: Oxford University Press.

White, A. (1971). *Inference. Philosophical Quarterly*, *21*(85), 289–302.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, *19*, 445–459.

# 7 Reasoning and Mental Action

*Markos Valaris*

## 7.1 Introduction

The phenomenon of reasoning, many philosophers would agree, is a paradigmatic case of *mental action* – an example of the active control rational creatures like us have over our own minds.[1]

The aim of this paper is to explain how this could be so. As it turns out, this is far from straightforward.

Some of the challenges to the idea that reasoning is an exercise of mental agency are familiar, and have been discussed before (if in somewhat different contexts): if I believe *R* and see that *R* conclusively supports *P*, then it seems like I am not at liberty as to whether to infer *P*: *P* is, for me, irresistible. Conversely, if I believe *R* and see no connection between *R* and *P*, I *cannot* infer *P* from *R*, no matter how hard I try. So, if inferring is an action, it must be a kind of action that we cannot simply *choose* to perform regardless of what else we believe.[2] For present purposes, I will set this issue aside. It is not, after all, entirely clear why we should think that all of our actions must be such that we can just choose to perform them regardless of what else we believe.

However, there is another, and arguably deeper, concern about how reasoning can be a mental action. This has to do with the metaphysics of reasoning. Philosophers writing on reasoning tend to take the phenomenon they study to combine two essential characteristics. On the one hand, reasoning is supposed to be an exercise of the control we have over our own cognitive states – paradigmatically, reasoning is a case of adopting or revising an attitude for reasons that the agent recognizes as such. On the other hand, reasoning is supposed to be a *process* of adopting or revising one's attitudes, where by 'process' I mean an occurrence that *develops* or *unfolds* over time. This combination of views is very widespread in the literature, and indeed frames the way the topic is generally introduced. To pick just two prominent recent examples, here is Paul Boghossian (2014, p. 2):

DOI: 10.4324/9780429022579-8

> [By 'inference' I mean] the sort of 'reasoned change in view' that Harman (1986) discusses, in which you start off with some beliefs and then, after a process of reasoning, end up either adding some new beliefs, or giving up some old beliefs, or both.

And here is John Broome (2013, p. 234):

> Active reasoning is a particular sort of process by which conscious premise-attitudes cause you to acquire a conclusion-attitude.

Similar views are endorsed by many others. Such claims, however, turn out to be misleading. There is no *single* mental phenomenon that meets both of the constraints these authors suggest. While we are capable of holding and revising attitudes for reasons, this is not usefully thought of as any kind of *process*. And, while deliberation clearly involves processes (for example, seeking out evidence and working out relations of support) these are not, as such, cases of adopting or revising any attitude: you can work out that one thing follows from another, for instance, without any inclination to adopt either of them.

I argue for this distinction at some length elsewhere (Valaris, 2018).[3] In Section 7.2 below I briefly revisit the issue and offer some new considerations in support. But the bulk of this chapter is devoted to a question downstream from that – namely, how do we reconcile this apparent fact about the metaphysics of deliberation with the idea that reasoning is a mental action?

The puzzle is this. On the one hand, we are accustomed to thinking of actions as events or processes, that is, as occurrences that, in some sense, *unfold* or *develop* over time.[4] This, then, would suggest that, when seeking to argue that reasoning is a mental action, we should look at the *processual* aspects of deliberation. In particular, we might settle on the activity of working out what follows from what, or what I will call 'deduction' in what follows. Since, however, working out what follows from what is not, as such, to adopt or revise any particular (categorical) attitude, settling the question of mental agency in this way would require giving up on the idea that mental agency in deliberation consists in exercising control over our own attitudes. What, then, would it take to recognize a form of mental agency that does consist in exercising control over our own attitudes? I will close with some rather tentative remarks on this topic.

## 7.2 The structure of deliberation

The problem I address in this paper arises because of the distinction I sketched above, between two aspects or components of deliberation. Consider an agent who notices the following:

1 Mandy's car is in the driveway.

Our agent, then, recalls her standing belief that:

2 Either Mandy is at home, or her car is not in the driveway.

Performing a trivial inference, the agent ends up believing:

3 Mandy is at home.

Our question is how to analyze episodes such as this one (call them episodes of deliberation): in particular, what cognitive operations should we ascribe to the agent?

I suggest that deliberation involves (at least) *two* distinct operations. On the one hand, the agent considers the contents of some of her existing beliefs – namely, that Mandy's car is in the driveway, and that either Mandy is at home or her car is not in the driveway – and notices that something further follows from them – namely, that Mandy is at home. As a result of her reflections, then, the agent is now in a position to *judge* or *believe* that Mandy is at home. And this seems to be a distinct operation.

Why should we think that these are two distinct operations? To begin with, notice that the cognitive capacities the agent exercises in each simply appear to be distinct. In the first case, the agent is exercising her capacity for *deduction* – for working out what follows from what. In the second, she is exercising her capacity for *believing something for reasons*. These are distinct capacities, which can be exercised independently of each other. To see this, notice that you have no trouble *agreeing* with our agent that (3) follows from (1) and (2) above. It would, however, be quite a stretch to suggest that you have *beliefs* regarding the whereabouts of (the entirely made up) Mandy. And the same goes, of course, for cases where we employ our deductive capacities towards purely schematic formulas, as in logic: you can deduce *B* from *A* and 'if *A* then *B*', even though *A* and *B* have been assigned no content, and so would not seem to be appropriate objects for belief or other propositional attitudes at all.[5]

Furthermore, it is a familiar point that working out that a certain conclusion follows from your premises does not, by itself, suffice for accepting that conclusion. It may, on the contrary, spur you to reject one or more of your premises. Working out what follows from what is one thing; making up your mind what to believe is quite another. The former is what I call 'deducing', and the latter 'reasoning' or 'inferring.'

I will discuss deduction in some more detail in the next section, but before getting to that let me clarify a couple of points. Deducing is clearly a *process*, in the sense that it is a type of occurrence that unfolds in time.

More precisely, using the typology familiar from Vendler (1957), it is an 'accomplishment' – that is, a process that leads towards a more-or-less determinate end-point. For example, if an agent (say, a student sitting a logic exam) deduced *P* from *R*, then there was a time *t* such that the agent *finished* deducing *P* from *R* at *t*, and a period of time leading up to *t* during which she was *deducing P* from *R*, but had not yet deduced *P* from *R*.[6] Furthermore, it is very plausible that the end-point of deduction – the state an agent performing a deduction ends up in, upon successfully completing it – is some kind of conditional doxastic attitude, which we could express in words as follows: '*P*, given *R*.' Such attitudes should not be confused with unconditional beliefs in (say) truth-functional conditionals. As understood here, a conditional belief expressible as '*P*, given *R*' is an attitude to the *pair* of propositions *R* and *P*, to the effect that, given *R*, no ways for things to be in which *P* fails to be true are possible.[7] Competently deducing *P* from *R* can *justify* you in holding such a conditional belief.

Let us, now, turn to the second of the two operations distinguished above, namely inferring. Inferring, in our sense, is what Boghossian calls 'reasoned change in view', that is, an exercise of our capacity for controlling our own cognitive states. The main contention that I want to defend in the remainder of this section is that inferring, understood in this sense, is *not* a process. The simplest argument for this claim is a negative one: as we saw, the process involved in episodes of reasoning is the process of deduction, and deducing *P* from *R* does not involve believing, or taking any other (unconditional) attitude towards *P*. So, *a fortiori*, it does not involve *changing* your view regarding *P*. Now, this argument may not appear conclusive: perhaps, even if inferring is not the same as deducing, there is some other process involved in episodes of deliberation, and it is *that* process which ought to be identified with inferring. But there is also a more principled reason to resist identifying inferring with any process at all.

I think that an essential feature of reasoning (at least active reasoning of the sort we are interested in here) is that it satisfies what has come to be called, due to the influence of Boghossian (2014), the 'Taking Condition':

> Inferring *P* from a set of premises *R* requires taking it that *R* supports *P* (in some suitable sense), and believing *P* in part because of this.

The condition as formulated focuses specifically on the case of reasoning from and to beliefs, but analogous claims would hold for reasoning to attitudes other than belief, such as intentions. I believe that the main reason for denying that inferring is a process is that it gives us the best chance of maintaining the Taking Condition. But let us begin by

considering why we should want to hold on to the Taking Condition in the first place.

The Taking Condition is attractive, because it captures the idea that reasoning is an exercise of our rational and reflective capacities. But there is also a more direct reason to suspect that the Taking Condition, and indeed a specific *version* of the Taking Condition, is true. This is because of the susceptibility of reasoning to a certain kind of defeat. In particular, it seems possible that an agent may be justified in believing some propositions *R* which support *P*, infer *P* from *R*, and yet *fail* to be justified in believing *P*, because she was not justified in *taking it that R* supports *P*.

To illustrate, suppose that a detective possesses well-confirmed evidence *E* (consisting, I am assuming, in a set of true propositions or facts), which, as it happens, conclusively supports the claim that Alma was guilty of a certain crime. Suppose, moreover, that our detective *infers* from this evidence that Alma is guilty. Still, it seems entirely possible that our detective has become convinced that *E* conclusively supports Alma's guilt for entirely the wrong reasons, or even no reasons at all (perhaps the detective's taking that this piece of evidence points to Alma's guilt is simply based on prejudice, for example). In such a case, it seems clear that the detective's belief that Alma is guilty is not justified. But, *why* is it not justified? Importantly, it looks like we cannot explain this by taking the detective to lack justification to believe any *premises* of her reasoning: after all, *prima facie* at least it looks like *E* comprises all of his premises, and, by hypothesis, the detective is justified in believing *E*.[8] The most plausible explanation why the detective is not justified in believing that Alma is guilty, I think, is that the detective is not justified in *taking it that E* supports the claim Alma is guilty.

There are two lessons that the possibility of defeat of this sort holds for us. First, since the detective's case is not far-fetched or even unusual, it suggests that the Taking Condition is true: inference *always* involves taking it that your premises support your conclusion. Second, it suggests that the 'takings' in question must be the sorts of thing that can be, or fail to be, justified. They must be states of a sort that admits of epistemic assessment. This suggests that they are *doxastic* states – beliefs of some sort. This does not mean that they must be *second-order* beliefs, about the agent's own beliefs: they can be entirely first-order conditional beliefs, expressible in words as '*P*, given *R*.' These, of course – and not coincidentally – are just the sorts of beliefs that deduction was argued to yield earlier.

I have, so far, explained why I think that inference is subject to the Taking Condition, and in particular a doxastic version of the Taking Condition. Familiarly, however, this is a position that many have thought untenable. Usually, this is because they have thought that regress arguments, going back to Lewis Carroll's (1895) "What the Tortoise Said to

Achilles", have shown that the Taking Condition would make inferring impossible (see, e.g., Boghossian, 2003; Brewer, 1995; Railton, 2006; Winters, 1983). I agree that Carroll's regress argument teaches us something important about the nature of inference. I do not think, however, that what it teaches us is that inference is not subject to the Taking Condition. What the regress argument shows, instead, is that there is no such thing as *a process* of inferring.

Carroll's own text is cryptic, and different authors have offered different interpretations of the regress argument. I think, however, that we can put the main idea as follows. As everyone would agree, merely believing the premises of an argument is not enough to infer its conclusion: reasoning surely requires more than that. But, what more? According to those who support the Taking Condition, reasoning always requires taking the premises of your argument to support its conclusion. Now, the core move in the regress argument is to insist that simply adding such takings is *still* not sufficient for inferring *P* from *R*. After all, such takings would seem to be – especially on the doxastic construal of the Taking Condition, which I defended above – simply more beliefs. In that case, they would appear to belong among the *starting points* of reasoning – that is, among your premise-beliefs:

> Any beliefs involved in reasoning (and which are not identical with your conclusion-belief) must be among your premise-beliefs.

It is easy to see how this principle, together with the Taking Condition, leads to a vicious infinite regress. According to the Taking Condition, to infer *P* from *R* you must take it that *R* supports *P*. But then, from the principle just stated, this further belief must be among your premise-beliefs. As a result, your reasoning to *P* does not proceed from *R alone*, but from *R* and the proposition that *P* follows from *R*. Applying the Taking Condition again, we must then take you to also believe that *P* follows from *R and* the proposition that *P* follows from *R*. Since the same reasoning can now apply all over again, we are facing an infinite regress of ever-increasing premise sets and complex connecting beliefs.

It seems clear, then, that friends of the Taking Condition need to reject the principle above. There must be a role for beliefs about how your premises connect to your conclusion in reasoning, which does not consign them among your premise-beliefs. But what could that role be?

As I have argued elsewhere, (Valaris, 2014, 2016a, 2016b), the best way to avoid this type of regress while holding on to the Taking Condition involves giving those beliefs a *constitutive* role in reasoning, in the following sense: inferring *P* from *R* is not a process that takes you *from R to P*, but simply your *believing P, by* believing *R* and recognizing that *P* follows from it.[9] As Alan White (1971, p. 291) puts it: "inference is not the passage from A to B, but the taking of B as a result of reflection

on A." (By 'the taking of B' I assume that White means something like 'taking *up of B* as a belief.') For example, if you believe that it rained last night, and that if it rained last night then the streets will be wet, then your inferring that the streets will be wet from these premises just is your *believing that* the streets will be wet, in a particular *way* – namely, *by reflecting upon* your premises, and recognizing that they support just this conclusion. If this is correct, inferring is not a process, but a *state* – a way of believing. For example, to infer that Mandy is at home from the claims that Mandy's car is in the driveway and that either Mandy's car is not in the driveway or she is at home just is to believe that Mandy is at home, by recognizing that it follows from two other things that you believe.[10]

So far as I can see, this view provides the cleanest and best motivated way to avoid the regress argument. It does so by finding a clear role in reasoning for the beliefs implied by the Taking Condition, without requiring them to figure among the reasoner's premise-beliefs.

Let us now put all of the pieces together. Return to our earlier example, of an agent reasoning from (1) and (2) to (3). As I suggested above, we should regard the agent as doing *two* things here, not one. The agent begins with two premise-beliefs, with (1) and (2) as their contents, respectively. According to the Taking Condition, this is not enough for her to infer (3): she also needs to recognize that (3) follows from (1) and (2). This, as we already saw, is a job for the agent's deductive capacities: the agent can *deduce* that, given (1) and (2), (3) must be the case. Since, as we may assume, our agent has not in the meantime acquired any new evidence against (1) and (2), she is now in a position to infer (3) – that is, to believe (3), by recognizing that it follows from (1) and (2). This, I take it, is paradigmatic of how deduction and inference work together in a simple episode of deliberation.

This, now, leads to our puzzle. If episodes of deliberation should be analyzed into two distinct operations, then what happens to our idea that reasoning is a mental action? Are both of these operations mental actions, or just one of them? And, relatedly, can we make of the idea that inferring – understood as a *way of believing* – is an action? I will consider these questions in turn.

## 7.3 Deduction, attention, and mental action

My aim in this section is to explain why I think deduction – the operation of working out what follows from what – is properly thought of as a mental action.

Part of the difficulty here has to do with the idea of mental action itself. In particular, there are no uncontroversial criteria on the basis of which to judge that something is, or is not, a mental action.

According to standard approaches in the theory of action, a bodily movement counts as an (intentional) action just in case it is caused, in

the right way, by suitable mental states of the agent – where these may include beliefs, desires, or intentions. Such mental states are supposed to simultaneously *cause* and *rationalize* the ensuing intentional actions. (The *locus classicus* for this view is Davidson (1963); for a sophisticated recent defense against objections, see Smith (2010).) But, as several authors have noted, it is doubtful that this style of approach will work here (see, in particular, Strawson (2003) and Gibbons (2009)).[11]

The core difficulty is this. Consider the *content* of mental states that are suitable for the double role of causing and rationalizing a given action, say an instance of *F*-ing. Notice, now, that – as standard action-theorists are well aware – the explanations that determine whether a given set of bodily movements counts as an action or not are *hyperintensional*: whether some mental states rationalize your *F*-ing *qua F-ing* depends on *how*, or under what description, they represent your *F*-ing. For example, even if your bodily movements somehow end up making an omelet, if you had no beliefs, desires, or intentions that represented your bodily movements *as* a way for you to make an omelet, then standard theories of action would have to conclude that you did not intentionally make an omelet: your omelet making was, at best, a side-effect of something else that you did. More generally, it seems that if your *F*-ing is to count, as such, as an action of yours, it needs to be explained in terms of attitudes that represent it *as* an instance of *F*-ing.[12]

The problem, now, is that this model seems inapplicable to the case of mental actions (or, at least, to large classes of them). Consider the case presently under consideration, namely that of deduction. What mental states would cause and rationalize your deducing *P* from *R*? In order to meet the above constraint, such mental states would need to represent this deduction *as* an instance of deducing *P* from *R*. But it seems very unlikely that our deductions are generally caused by such states. While sometimes (as in a logic exam) we really do set out to deduce a specific conclusion from a set of premises, for the most part, it would seem that we do not. After all, if I already *know* or *believe* that *P* follows from *R*, then why go through the trouble of deducing it again?

A natural move at this point would be to relax the above constraint. Perhaps a mental event can count as a mental action in virtue of its causes, even if those causes do not represent that event in full detail – specifically, even if they do not fully fix its content. In particular, perhaps we might suggest that deducing *P* from *R* can count as an action if caused by a desire (or intention) simply to see *what follows from R*, rather than deducing specifically *P* from *R*.[13] Unfortunately, however, this proposal seems to be too weak for our purposes. In particular, it would seem to classify as mental actions certain processes that, intuitively, are not such.

Mele's (2009) case study of remembering makes this point very well. As Mele points out, we can actively decide to search our memory for a

piece of information that we cannot currently recall. But, as in the case of judgment, our relevant intention cannot be to search our memory *for that specific piece of information*, under that description. Suppose, for example, that I have forgotten someone's name, and I need to search my memory for it. Clearly, I cannot decide to search my memory specifically for the information that this person is Joe Smith; rather, I need to search my memory for this person's name, whatever it may turn out to be. So, we need to ask, does this suffice to make my memory search a mental action? I think we ought to agree that it does not. I *initiate* the memory search and, if all goes well, a few seconds or minutes later the name just pops up in my consciousness. However, and crucially, while the process is ongoing it seems to require no input or intervention of any kind from me. Indeed, avoiding focusing on the question at hand too directly sometimes seems to improve our chances of success at a task of recall. For example, if simply asking myself what the name of a particular person proves to be ineffective, it may be better to focus instead on different matters, such as reconstructing my most recent interactions with him, hoping that this will 'prime' my memory in the right way (Mele, 2009, p. 19).[14]

Using Mele's (2009) terminology, we should distinguish between my *bringing it about that* I remember the person's name, and my *remembering* the person's name. The former may well be a mental action – specifically, the mental action of *initiating* a memory search that ultimately succeeds. The latter, however, should either be identified with the sub-personal and non-agentive processes that actually search my memory, or with their outcome. In the latter sense, we might say that remembering the person's name was something that I *intentionally did*; but, I take it, what this means is simply that it was the intended outcome of an action of mine. Switching authors and terms, we might say that active remembering fits the model of 'mental ballistics' that Strawson (2003) advocates as a general, and explicitly deflationary, account for mental agency: in active remembering, we jog our memory into action, but then we simply have to let relevant sub-personal processes take their course.

Deduction, however, is not plausibly ballistic in this way. For example, while deducing, you need to keep your mind focused on the task at hand; turning your attention to other matters, in the hopes that this will somehow 'prime' your deductive capacities, is not going to improve your chances of getting it right. This suggests that what makes us think of deduction as a mental action is not its initiating conditions. It is, rather, that the process of deduction is *guided* or *controlled* by us, for its duration.[15]

But what exactly does this mean? What sort of process *is* deduction, and how is it guided by us? The nature of agential control or guidance, in general, is a matter of debate in the philosophy of action, and I cannot discuss it here.[16] All I can do is simply explain how I am thinking of the process of deduction, and the sense in which I take it to be guided by us.

The central component of the account I want to suggest is *intellectual attention*. By calling this type of attention 'intellectual', I do not mean to suggest that deduction involves attending *to* your own mental states or processes. Return to our example from the previous section, of an agent who works through the following argument:

1 Mandy's car is in the driveway.
2 Either Mandy is at home, or her car is not in the driveway.

Therefore:

3 Mandy is at home.

I think a natural way to reconstruct the agent's thinking might be as follows.

> Our agent considers Mandy's whereabouts, and specifically whether she is at home or not. She then notices that the whereabouts of Mandy's car are relevant to whether Mandy is at home or not. More specifically, she notices that there are two possibilities relevant to her concerns: either Mandy is at home, or else her car is not in the driveway (premise (2)). But Mandy's car *is* in the driveway, ruling out the latter possibility (premise (1)). So, given premises (1) and (2), Mandy must be at home.

For present purposes, the thing to note about the above reconstruction is that what the agent attends to is not her own thinking.[17] Rather, the agent considers various possibilities, or ways for things to be, and their relations to each other. More specifically, the agent considers ways for her conclusion to be false, until she satisfies herself that no such ways are consistent with her premises. She thus rules out ways for things to be in which (1) and (2) are true while (3) fails to be true.

But if the agent is not attending *to* her own thinking, then in what sense is her thinking controlled by her? I think the answer here lies in the way in which, as thinkers, we can direct *our attention to* this or that subject-matter. Returning to our example, the agent must make sure she directs her attention to the question at hand – specifically, the question of Mandy's whereabouts, given (1) and (2). Then, she needs to make a note of what possibilities she needs to consider, and then attend to each of them in turn, in light of her premises.

There is nothing unfamiliar, I take it, about the appeals to directed attention mentioned in the previous paragraph. Still, it would be nice if we could give something more by way of a theoretical account of our capacity for directed attention. While I cannot give anything like a full account, I can at least give some pointers to how I am thinking of it.

In keeping with other recent authors, I take it that attention is not a 'first-order' cognitive or processing resource. Rather, attention concerns the *allocation of other* cognitive resources.[18] In the case of *perceptual* attention, attending to something is primarily a matter of directing your *perceptual* resources to it. In the case of intellectual attention, it is a matter of directing your *intellectual* resources to one subject-matter rather than another. When it comes to deduction, in particular, the relevant resource is our capacity to consider possibilities.

Unfortunately, I have no full analysis of the notion of considering possibilities either. Still, I take it that this notion is intuitively familiar. For example, if you consider the question, whether Mandy is at home, then you are *ipso facto* considering two mutually exclusive ways for things to be, namely, her being at home and her not being at home.

Furthermore, although I described the above as a reconstruction of our agent's thinking, I do not mean to suggest that it is a transcript of phenomenally conscious inner speech that the agent must have engaged in. It seems plausible, for example, that deductions as simple as the one illustrated above can be performed by a typical agent much faster than it would take them to explicitly verbalize all relevant considerations in inner speech. My reconstruction aims to capture the thoughts that play a role in our agent's deduction, regardless of how – and even whether – these thoughts figure explicitly in her stream of consciousness at the time.

It may seem surprising that a reconstruction of an agent's personal level thought should appeal to thoughts that are not, at their time of occurrence, phenomenally conscious as inner speech.[19] However, it should not. When crossing a busy street at a crosswalk, a lot of thoughts regarding traffic conditions must, in some way, play a role in rationally shaping your behavior. A faithful account of how you succeed in crossing the road would surely need to mention those thoughts. And yet this does not mean that you need to *consciously verbalize* all these thoughts – indeed, if you were to attempt to do so, you would likely undermine your chances of getting across the street safely and quickly. Similarly, my reconstruction of our agent's thinking is meant to capture the rational or epistemic content of her deduction, not necessarily its phenomenal character. For example, all that goes on in our agent's stream of consciousness might be this. She asks herself, 'Is Mandy at home?'; with that question still in mind, she then notices the car in the driveway; and straightaway, she answers her own question in the affirmative.

Noting the distinction between the epistemic content and the phenomenal character of deduction has a further consequence. We are accustomed to thinking of deductions in linguistic terms, just as above. And, indeed, when we choose to focus our attention on our own deductions (something that, as argued above, is not necessary for deduction as such) what we end up introspecting typically seems to take the form of inner speech. But I do not think we can conclude from this anything about

the nature of the representational vehicles of the thoughts constitutive of our deductions. *We* use words to express our deductions, even to ourselves; but it does not follow that the cognitive mechanisms that underlie our deductive capacities must do so as well. This, I am assuming, is an empirical question, which cannot be settled simply by introspection.[20]

This, I hope, gives us a better grip on the sense in which deducing is a mental action. Let me now turn to the question of whether inferring – understood, as suggested above, as a *way of believing* – is a mental action.

## 7.4 Inferring as a mental action

Recall that, as I argued in Section 7.2, while *deducing* should be thought of as a process, *inferring* should not. Deducing, if all goes well, results in a conditional belief, to the effect that something holds, given something else. This is not, as yet, to draw any categorical conclusions about anything. *Inferring* one thing from another, by contrast, implies *believing* the former thing, in virtue of believing the latter and recognizing a connection between them. In this sense, inferring is not a process, but a way of believing. Is there, then, any way in which inferring can be understood as *active*?

Let us begin by sharpening our intuitions in this area. If inferring is a kind of believing, then how does it differ from *other* kinds of believing? Consider an ordinary standing belief, such as my belief that Napoleon lost the battle of Waterloo. This, it seems, is a rational belief (indeed, plausibly, a piece of knowledge). But what are my *reasons for* this belief? The answer to this question is far from clear. On the one hand, one might try to track the origins of this belief. Presumably, my belief that Napoleon lost the battle of Waterloo originated in some long-forgotten history lesson, or perhaps a history book or television program. Either way, my belief presumably originated through testimony by some trustworthy authority. The relevance of this, however, is doubtful. While my receipt of such testimony may have been my reason for believing that Napoleon lost the battle of Waterloo at one time, it is far from clear that this can be my reason for believing that Napoleon lost the battle of Waterloo *now* – how could it be, if I have completely forgotten all about it? This sort of problem has led epistemologists to look for alternative accounts for the rationality of standing beliefs. For example, we may think that, in the absence of substantial conflict with other things that I believe, my belief retains its rationality by some sort of default; or we may think of *remembering that* as a source (or form) of knowledge itself.[21]

For present purposes, I do not need to defend an account of the rationality of such standing beliefs. The crucial point is just that, whatever its rational credentials, my belief that Napoleon lost the battle of Waterloo is not something that I *infer* from anything else; I do not

believe that Napoleon lost the battle of Waterloo as a result of reflecting on anything else I believe. Even if we adopt an account according to which the present rationality of my belief derives from its past reasons, it is surely not plausible that I *now* hold the belief by *reflecting upon* those reasons: reasons I have forgotten about are not available to me to reflect upon.

This obviously contrasts with cases of inference. If the agent in our earlier example infers (3) from (1) and (2), then she now *believes that* Mandy is at home *by* taking it that Mandy's being at home follows from other things that she believes. In a case like this, there is no mystery regarding the agent's *reasons for* believing that Mandy is at home. For instance, if asked, our agent would plausibly not hesitate to supply (1) and (2) (or close paraphrases) as her reasons for believing that Mandy is at home. In contrast to my belief that Napoleon lost the battle of Waterloo, this is a case in which the agent believes something as a result of reflecting upon other things which she believes.[22]

It should be obvious that, while many of our beliefs start out as inferences, most of them quickly transition to being merely standing beliefs. It is not hard to see why. Believing that *P* by reflection on *R* is clearly a rather demanding way of believing: you need to be thinking not just about *P*, but also about *R* and the connection between those thoughts, in a way such that the former thoughts are settled by the latter. Forging and maintaining such connections in thought is not an effortless matter. Absent special reasons for doing so, we are normally (and reasonably) not inclined to devote the resources necessary for it. For example, on the plausible assumption that our agent cares more about Mandy's whereabouts than those of Mandy's car, she may soon forget *how* she initially came to believe that Mandy is at home, while retaining the latter belief. If one of our beliefs is challenged (or its grounds become relevant for other reasons), then it may be possible for us to *rediscover* the grounds on which we first held it. But this need not always be possible, and even when it is, it will amount to a *new* episode of inferring – even if the contents stay the same.

So, what does the contrast between inference and standing beliefs tell us about the sense in which inferring is a mental action? The label itself does not, of course, matter very much. It is worth noting, however, that this type of belief, unlike the case of a mere standing belief, involves an exercise of *control*. This does not mean that it is a case of believing at will. For instance, the agent in our earlier example quite plausibly *could not* – whether as a matter of metaphysical or merely psychological impossibility – simply choose to believe that Mandy is not at home, given her other beliefs and perceptual evidence. Nevertheless, she exercises control in a more subtle way: she ensures that her view of one thing (namely, Mandy's whereabouts) is *settled by her view of something else*, which is more directly knowable (the whereabouts of Mandy's car).[23]

In this sense, the control we exercise in inference is similar to the control over our sensory states we exercise in active perception, for example in *looking*, *listening*, or *watching*. Suppose I watch a bird fly over my backyard. The reason why we are inclined to count watching as *active* rather than passive, I suggest, is this: it involves exercising a certain kind of control over our visual experiences. But this does not mean that in watching the bird I can freely *choose* what I perceive. What my control consists in, rather, is my ensuring that my perceptions – whatever their content may turn out to be – derive from a particular source, namely, the bird as it flies over my backyard.[24] Similarly, when I infer *P* from *R*, my control over my beliefs does not consist in freely choosing to accept *P*, but rather in ensuring that what I think about *P* is determined by a particular set of considerations, namely *R*.

This point should also help moderate the weight of an *ontological* objection against treating inference (in the present sense) as a mental action. In particular, as I suggested earlier, we are accustomed to thinking about actions as events or processes, that is, as occurrences that (in some sense) *unfold* over time. By contrast, in my view inferring is a *way of believing*. As such, it would seem to be a state, not an event or process; and, one might worry, we cannot make sense of a state being an action. For example, Vendler (1957, p. 148) remarks that states "cannot be qualified as actions at all", after observing (ibid., p. 144) that you cannot answer the question 'what are you doing?' by saying 'I am knowing (or loving, recognizing, and so on).'

I think, however, that the fact that we count things like listening, looking, and watching as actions should lead us to doubt the force of this objection. Now, it is true that on a standard Vendlerian classification, perceptual acts like these would count as *activities* or *processes* rather than states. Thus, you could, for example, answer the question 'what are you doing?' by saying 'I am watching a bird fly over my back yard.' It is not, however, clear that this difference should be determinative.[25] After all, as I have suggested, in both cases what you do consists of exercising a certain sort of control over some of your own mental states, by allowing them to be determined in a certain way. If we count this as an exercise of agency or control in the perceptual case, then perhaps we should also count it as such in the doxastic case as well.

How does the view just sketched compare to existing accounts of doxastic agency? I only have space here to compare it with one such view, namely, the one proposed by Pamela Hieronymi (2009). Hieronymi's discussion is framed around the idea that some of our attitudes – paradigmatically, core cases of beliefs and intentions – embody our answers to questions. (For example, believing that the grass is green embodies your affirmative answer to the question, 'is the grass green?') I cannot here discuss this proposal with any generality. For present purposes, the

important point is that Hieronymi argues that attitudes that embody answers in this way can also be understood as active.[26]

Hieronymi understands agency in terms of control (just as I do here). She thus argues that there is a distinctive kind of control we exercise over our beliefs – what she calls *evaluative control*. She introduces this notion as follows:

> We might say that we control those aspects of our minds [*viz*. beliefs, and other propositional attitudes] because, as we change our mind, our mind changes – as we form or revise our take on things, we form or revise our attitudes. I call this exercising *evaluative control* over the attitude.
>
> (2009, p. 140)

This passage is meant to help us understand what sort of control you have over your doxastic attitudes. It involves, however, a crucial ambiguity. In particular, according to this passage, our control over our doxastic attitudes is supposed to be explained in terms of our 'forming or revising our take on things.' But which things are these? In particular, is the subject-matter of those takings the same or different from the subject-matter of the doxastic attitudes in question? Suppose, first, that they are not meant to be different: your control over your doxastic attitude to some proposition *P* is supposed to be explained in terms of your take on *P* itself. If this is the view, however, then this passage does not succeed in identifying a meaningful kind of control, for it does not tell us how you can form or revise your take on *P* otherwise than by forming or revising your doxastic attitudes to *P*. Hieronymi's claim threatens to collapse to the triviality that you can form or revise your attitudes by forming or revising your attitudes.

Things look better, however, if we read the passage as suggesting that you can control your doxastic attitudes regarding a certain proposition *P* *through your take on some other propositions*. In this case, it seems very plausible that the phenomenon that Hieronymi has in mind is just what I have been discussing under the heading of 'inference', that is, believing one thing by reflecting upon another. For example, our agent's exercising evaluative control over her beliefs regarding Mandy's whereabouts would consist in letting her doxastic state on *this* topic be settled by her beliefs on something else – namely, the whereabouts of Mandy's car, and the general connection between the car's whereabouts and Mandy's. Thus, assuming that this is the right way to understand Hieronymi's notion of evaluative control, inference is a case – perhaps the archetypical case – of evaluative control.

Let me close by briefly taking stock of the argument of this paper. I began by noting that it is natural to think of reasoning or deliberation as a mental *action* on our part, and in particular as an exercise of a

distinctive kind of control that rational creatures like ourselves have over our own minds. But, while this idea has widespread support, it is not straightforward to make good on. As I argued, to make progress on this project we must begin by noting that deliberation is, in fact, a complex process, involving at least two distinct operations – namely, what I called 'deducing', and what I called 'inferring.' As we have seen, there are reasons to think of both of these as exercises of agency or of control – even if in rather different ways.

## Notes

1. Of course, terms such as 'reasoning' or 'inference' are sometimes used to refer to things that are clearly not personal-level mental actions, such as the information processing that occurs in early vision. Such instances of so-called 'unconscious inference' are not our topic here. Less clear-cut are cases of so-called 'System 1' reasoning, such as the seemingly automatic, intuitive and heuristic-based judgments elicited by numerous famous psychological experiments. On some views, such inferences are outside the sphere of rational control, and so presumably ought not to count as mental actions either. While I do not think that this interpretation of the data is mandatory (for discussion of this question, see the essays in Evans and Frankish (2009)) and thus I believe that some instances of System 1 reasoning could be construed as plausible candidates for the status of mental action, I will generally focus on cases that are more clearly personal-level and reflective.
2. This issue is generally discussed under the heading of 'believing at will' (a topic that has been on our radar since Williams (1973)), but of course just the same issues arise with regard to inferring at will.
3. This distinction is not commonly drawn in the recent literature. Rumfitt (2011, 2015) draws a similar distinction, but does not elaborate on the nature of what he – like me – calls 'inference.' An older generation of authors were more prone to drawing distinctions in this area, though it is not in every case easy to see how their distinctions map onto the distinction I am drawing here. Ryle ([1945] 2002) argues that reasoning is not a process, but in doing so seems to miss the existence of what I call 'deduction.' White (1971) seems to come close to the distinction I am after, as we will see below.
4. Vendler (1957) introduced a four-fold classification of verbs into 'accomplishments', 'achievements', 'activities', and 'states.' While Vendler described his classification scheme in linguistic terms (as a four-fold classification of lexical verbs), many have thought that it has deeper ontological significance, and in particular that it marks different ways of occupying *time* (Mourelatos, 1978). Accomplishments and activities are dynamic occurrences that unfold or develop over time, and are expressed by verbs and verb-phrases that admit of progressive aspect – such as *walking to the store*, *running*, or *thinking*. Achievements and states, on the other hand, are not supposed to unfold over time, although for very different reasons: achievements are non-durative events that mark the limits of processes (like *arriving* at your destination, or *starting* the race), while states are conditions that one is *in* for a period of time. For present purposes, all we are interested in is the apparent distinction between states (such as believing) and processes (such as working out a proof).

5. A number of philosophers have noted that many examples of what is standardly called 'reasoning' in the literature do not seem to involve reasoning from things that you believe to other things that you believe (see, e.g., (Balcerak Jackson & Balcerak Jackson, 2013; Wright, 2014; Dogramaci, 2016)). These authors, however, tend to characterize the phenomenon they notice as 'suppositional' reasoning, where supposition is meant to be some kind of attitude of acceptance, somewhat like a belief (perhaps a mock or pretend belief). Even this, however, seems wrong. As I argue in Valaris (2018), we can see this by considering the Moorean proposition expressed by 'if it is raining then I do not accept that it is raining, and it is raining.' You can clearly work out that from this it follows that I do not accept that it is raining. It seems unlikely, however, that you can rationally adopt any attitude of acceptance towards this proposition. As Soteriou (2013, pp. 261–262) points out (though for rather different reasons), supposing something for the sake of the argument does not seem to be a self-standing propositional attitude at all.
6. Ryle ([1945] 2002, pp. 301–302) claims that "'I began to deduce, but had no time to finish' is not the sort of thing that can be significantly said." To the extent that by 'deducing' we mean, as I am doing here, 'working out the consequences of a set of hypotheses', then it seems that Ryle is simply wrong: an agent can clearly run out of time to complete her deductions, as anyone who has ever had to mark logic exams can attest. I suspect that Ryle is guilty of running together deduction with reasoning, understood as the distinctive kind of control rational agents have over their cognitive states.
7. This characterization raises some difficult questions that I cannot go into here. In particular, notice that if *P* really does follow from *R*, then any non-empty 'ways for things to be' ruled out by a belief expressible by '*P*, given *R*' cannot be classical possible worlds. Thus, on a standard classical worlds picture of content, all such conditional beliefs would seem to be vacuous. This seems puzzling, since working out what follows from what often is a substantive cognitive achievement. I take it, however, that this is a problem for the classical possible worlds theory of content, not for the account of deduction sketched here. Wedgwood (2012) also emphasizes the importance of conditional beliefs, which he understands as states of *accepting* arguments. Wedgwood thinks of conditional beliefs in terms of conditional credences. Since, however, he allows that a (non-ideal) agent's credences may be probabilistically incoherent, he must also not be thinking of the *objects* of the agent's credences as sets of classical possible worlds. Again, it looks like a more fine-grained account of content is needed.
8. Some might argue that, despite appearances, the detective's reasoning also includes many more hidden premises, some of which he may lack justification to believe (such a view is suggested, for example, by Tucker (2010)). Such views, I think, ignore or distort a distinction crucial to the theory of reasoning, between premises and background assumptions. Background assumptions play a crucial role in reasoning, by delimiting the space of possibilities the reasoner needs to consider: for example, if I am planning my next summer vacation, I will typically not consider the possibility that the external world is an illusion spun by an evil demon. It would distort the nature of my reasoning, however, to include the claim 'the world is not an illusion spun by an evil demon' among its premises.
9. In Valaris (2014), I made a distinction between 'basic' and 'non-basic' reasoning and applied the schema in the text to non-basic reasoning only. The reason why I made this distinction was that I could not see where the

beliefs required by the Taking Condition could come from, if not from further reasoning – thereby leading to another regress. But I now think this distinction was a mistake, caused by my own failure to distinguish between *deduction* and *inference*. The beliefs required by the Taking Condition are conditional beliefs, and are in every case supplied (and justified) by *deduction*. Since deduction is not the same thing as inference, no regress threatens.

10. Neta (2013) develops an account of inference according to which inference is a judgment. His account is similar in some respects to mine. Neta, however, takes it as essential to inferring that the thinker have thoughts about her own thoughts. I doubt that this is so, but I cannot engage in this debate here.
11. Wright (2014) suggests that we treat reasoning as an intentional action. While Wright does not elaborate on what theory of intentional action he endorses, assuming he is endorsing something like the standard view he would have to deal with the problems outlined below.
12. This is reminiscent of what Bratman (1987) calls the 'simple view' of intentions. Bratman famously criticizes the simple view, but his objections do not affect the main point, since he still maintains that your intentions must represent your actions as *attempts to F*. Other authors (e.g., Harman, 1999) have noted that sometimes we may take agents to *F* intentionally, even though *F*-ing was merely a foreseen consequence of what they were doing, rather than an intended or desired one. However, even in such cases it remains the case that the agent's mental states must have represented her actions as *F*-ing.
13. This kind of move is familiar in the literature. For example, Peacocke (1999, pp. 209–210) suggests that what distinguishes *directed*, as opposed to *idle*, conscious thought is that it involves the "intention to think a thought which stands in a certain relation to other thoughts or contents." Since deduction (as I am using the term here) is a paradigmatic instance of directed thought, this is presumably what Peacocke would say about this case as well. Claims similar to Peacocke's are also to be found in O'Shaughnessy (2003, p. 221) and Shah and Velleman (2005).
14. Can you at least reliably *stop* a memory search that has already been initiated? Though this is ultimately a matter for empirical psychology, the answer seems to be 'no.' *Consciously trying* to avoid thinking of someone's name is clearly just as ineffective as consciously trying to fall asleep. Turning your attention to other matters can certainly cause the search to fail, but it is far from guaranteed to do so.
15. As Frankfurt (1978) argued, this may be the case for actions in general, not just deducing or even just mental action. This is an important question, but I cannot discuss it further here. Strawson (2003, pp. 231–232) seems to allow that thinking may be directed in this way, but he takes this to be consistent with its being ballistic. This, I think, stretches the metaphor to a breaking point. What *makes* a process ballistic, it would seem, is that after setting the initial conditions, I have to let it take its course – just as, when I throw a stone at a target, after the stone leaves my hand I have no further control over its trajectory.
16. Most of the discussion of agential control in the literature has focused exclusively on the case of bodily action (Fridland, 2014; Shepherd, 2013; Wu, 2011, 2016); for the application of some ideas arising from the neuroscience of motor control to the case of mental action, see Campbell (1999) and Proust (2009). While I cannot discuss this literature here, I want to flag that there are some approaches to control or guidance that would *not*

appear to rule out ballistic processes from counting as 'controlled.' In particular, on some views a process can count as controlled by an agent just in case the agent is sufficiently reliable, across a range of counterfactual situations, in getting the intended outcome (Shepherd, 2013). Such views, however, would count remembering as a process that is under your control (assuming your memory is reliable enough). This is not the notion of control we need here.

17. Peacocke (1999, pp. 208–211) argues that while in directed thinking your own thoughts are not the *objects* of your attention, they nonetheless *occupy* your attention (or at least can do so). I am not sure I understand Peacocke's notion of something's *occupying* your attention without being the *object of* an episode of attention. I certainly agree with part of Peacocke's view, which is that directed thinking can be phenomenally conscious, in the sense that it can be part of what it is like for you at a given time.
18. This, of course, leaves plenty of room for further disagreement. Mole (2010) suggests that attending to a task consists in committing all relevant and available resources to it. Watzl (2017), by contrast, suggests that attention is a matter of comparative priority in a priority structure that determines the allocation of resources. Such disagreements do not matter for our purposes.
19. A number of philosophers appeal to inner speech as partly constitutive of occurrent thought, or at least a very reliable introspectable 'symptom' of thinking activity (Byrne, 2010; Carruthers, 2011; O'Shaughnessy, 2003; Ryle, 2002; Soteriou, 2013). Inner speech *may* sometimes be intimately connected to occurrent thought, but we should not assume that it must be.
20. For example, the highly influential mental models theory of the psychology of deduction relies on a mixture of linguistic and imagistic representations (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991).
21. The first option is defended by, among others, Harman (1973) and Wedgwood (2012). The second would appear attractive to infallibilists and 'knowledge-first' epistemologists, such as McDowell (1995), Neta (2011) and Williamson (2000).
22. Wedgwood (2012) also argues for a theory of rationality that distinguishes between the rationality of standing beliefs and that of acts of inferring. My approach differs from Wedgwood's, however, because he thinks of acts of inferring as events or processes of belief-formation.
23. Boyle (2009) argues that merely *standing* beliefs are also exercises of agency, because they involve the same capacities for self-determination that are also exercised in cases of reflectively making up our minds. I need not deny this claim, or even a sense in which standing beliefs may count as 'active.' It still remains the case that in inferring you are actually *exercising* capacities for doxastic control that are merely latent in cases of standing belief.
24. For an account of the perceptual activity of watching broadly consistent with this, see Crowther (2009). On Crowther's view, the activity of watching consists in maintaining perceptual contact with an object, with a view to acquiring knowledge about it.
25. See also Boyle (2009, p. 137) on this.
26. More carefully, Hieronymi (2009, pp. 141–142, n. 4) distinguishes between two readings of the claim that some attitudes embody answers: on one reading, this requires having *settled on* the answer, after considering the question; on a different reading, it simply requires being *committed to* the answer, even if you have never actually considered the question. Despite expressing sympathy for the latter view, Hieronymi restricts her discussion to the former (as easier to defend). I follow her on this.

## References


Balcerak Jackson, M., & Balcerak Jackson, B. (2013). Reasoning as a source of justification. *Philosophical Studies*, *164*(1), 113–126.

Boghossian, P. (2003). Blind reasoning. *Aristotelian Society Supplementary Volumes*, *177*, 225–248.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1), 1–18.

Boyle, M. (2009). Active belief. *Canadian Journal of Philosophy*, *39*(sup1), 119–147.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Brewer, B. (1995). Mental causation II: Compulsion by reason. *Aristotelian Society Supplementary Volumes*, *69*, 237–253.

Broome, J. (2013). *Rationality through reasoning*. Chichester, UK: Wiley Blackwell.

Byrne, A. (2010). Knowing that I am thinking. In A. Hatzimoysis (Ed.), *Self-knowledge* (pp. 105–124). Oxford: Oxford University Press.

Campbell, J. (1999). Schizophrenia, the space of reasons, and thinking as a motor process. *The Monist*, *82*(4), 609–625.

Carroll, L. (1895). What the tortoise said to achilles. *Mind*, *4*, 278–280.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Crowther, T. (2009). Watching, sight, and the temporal shape of perceptual activity. *Philosophical Review*, *118*(1), 1–27.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.

Dogramaci, S. (2016). Reasoning without blinders: A reply to valaris. *Mind*, *125*(499), 889–893.

Evans, J., & Frankish, F. (Eds.). (2009). *Two minds: Dual processes and beyond*. Oxford: Oxford University Press.

Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly*, *15*(2), 157–162.

Fridland, E. (2014). They've lost control: Reflections on skill. *Synthese*, *191*(12), 2729–2750.

Gibbons, J. (2009). Reason in action. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 72–94). Oxford: Oxford University Press.

Harman, G. (1973). *Thought*. Princeton, NJ: Princeton University Press.

Harman, G. (1986) *Change in view*. Cambridge, MA: MIT Press.

Harman, G. (1999). Practical reasoning. In *Reasoning, meaning, and mind* (pp. 46–75). New York: Oxford University Press.

Hieronymi, P. (2009). Two kinds of agency. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 138–162). Oxford: Oxford University Press.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P., & Byrne, R. (1991). *Deduction*. Hove, UK; Hillsdale, USA: Psychology Press.

McDowell, J. (1995). Knowledge and the internal. *Philosophy and Phenomenological Research*, *55*, 877–893.

Mele, A. (2009). Mental action: A case study. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 17–37). Oxford: Oxford University Press.

Mole, C. (2010). *Attention is cognitive unison: An essay in philosophical psychology*. Oxford: Oxford University Press.
Mourelatos, A. P. D. (1978). Events, processes, and states. *Linguistics and Philosophy*, *2*(3), 415–434.
Neta, R. (2011). A refutation of cartesian fallibilism. *Noûs*, *45*(4), 658–95.
Neta, R. (2013). What is an inference? *Philosophical Issues*, *23*(1), 388–407.
O'Shaughnessy, B. (2003). *Consciousness and the world*. New York: Oxford University Press.
Peacocke, C. (1999). *Being known*. Oxford: Oxford University Press.
Proust, J. (2009). Is there a sense of agency for thought? In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 253–280). Oxford: Oxford University Press.
Railton, P. (2006). How to engage reason: The problem of regress. In R. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: Themes from the moral philosophy of Joseph Raz* (pp. 176–201). Oxford: Oxford University Press.
Rumfitt, I. (2011). Inference, deduction, logic. In J. Bengson, & M. A. Moffett (Eds.), *Knowing how: Essays on knowledge, mind, and action* (pp. 334–360). New York: Oxford University Press.
Rumfitt, I. (2015). *The boundary stones of thought: An essay in the philosophy of logic*. New York: Oxford University Press.
Ryle, G. (2002). *The concept of mind*. Chicago: University of Chicago Press.
Shah, N., & Velleman, D. (2005). Doxastic deliberation. *The Philosophical Review*, *114*(4), 497–534.
Shepherd, J. (2013). The contours of control. *Philosophical Studies*, *170*(3), 395–411. doi: Retrieved from https://doi.org/10.1007/s11098-013-0236-1.
Smith, M. (2010). The standard story of action. In J. H. Aguilar, & A. A. Buckareff (Eds.), *Causing human actions: New perspectives on the causal theory of action* (pp. 45–57). Cambridge, MA: MIT Press.
Soteriou, M. (2013). *The mind's construction*. Oxford: Oxford University Press.
Strawson, G. (2003). Mental ballistics or the involuntariness of spontaneity. *Proceedings of the Aristotelian Society*, *103*, 227–256.
Tucker, C. (2010). When transmission fails. *Philosophical Review*, *119*(4), 497–529.
Valaris, M. (2014). Reasoning and regress. *Mind*, *123*(489), 101–127.
Valaris, M. (2016a). What reasoning might be. *Synthese*, *194*, 2007–2024.
Valaris, M. (2016b). What the tortoise has to say about diachronic rationality. *Pacific Philosophical Quarterly*, *98*(S1), 293–307.
Valaris, M. (2018). Reasoning and deducing. *Mind*, *128*(511), 861–885.
Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*(2), 143–160.
Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford: Oxford University Press.
Wedgwood, R. (2012). Justified inference. *Synthese*, *189*(2), 273–295.
White, A. R. (1971). Inference. *The Philosophical Quarterly*, *21*(85), 289–302.
Williams, B. (1973). Deciding to believe. In *Problems of the self* (pp. 136–151). Cambridge, UK: Cambridge University Press.
Williamson, T. (2000). *Knowledge and its limits*. New York: Oxford University Press.
Winters, B. (1983). Inferring. *Philosophical Studies*, *44*, 201–220.

Wright, C. (2014). Comment on Paul Boghossian, 'What is inference.' *Philosophical Studies*, *169*(1), 27–37.

Wu, W. (2011). Confronting many-many problems: Attention and agentive control. *Noûs*, *45*(1), 50–76.

Wu, W. (2016). Experts and deviants: The story of agentive control. *Philosophy and Phenomenological Research*, *93*(1), 101–126.

# 8 Causal Modeling and the Efficacy of Action

*Holly Andersen*

## 8.1 Introduction

This paper brings together Thompson's naive action explanation with interventionist modeling of causal structure to show how they work together to produce causal models that go beyond current modeling capabilities. I will, in the process, show why the internal structure of action, where stages are unified by rationalizations into a coherent overarching action, cannot be causal. Actions, and action explanations, cannot be reduced or simplified to causation and mere causal explanation without genuine loss. Despite this, existing causal modeling techniques can be deployed to model action in some cases. By deploying well-justified assumptions about rationalization, we can strengthen existing causal modeling techniques' inferential power in cases where we take ourselves to be modeling causal systems that also involve actions. This capacity for naive action explanation to strengthen causal modeling inferences provides motivation to incorporate it into interventionist approaches to causation.

Action explanation and interventionism are, in many ways, an awkward fit. The former involves all the rich particularities of singular instances of action, rich with normative structure. The latter is built for general pre-specified variables with allowed values, lacking the rich normative structure that is distinctive of action. Unification might seem like a tempting motivation to accommodate (or more likely, offer a reduction of) action explanations within the ambit of causal explanation. But such a move would result in an under-description of genuine structure in the world. Action explanation cannot be reduced to or fully supplanted by causal explanation. And conversely, causal explanation can be better understood by contrasting it with the kind of structure Michael Thompson (2008) calls rationalization. Action explanations involve a modal strength connecting the relata that makes them much closer to what Lange (2012) has called distinctively mathematical explanations, rather than the modal strength by which causal relata are connected. Just as distinctively mathematical explanations cannot be reduced to any

DOI: 10.4324/9780429022579-9

collection of causal explanations, no matter how exhaustive, neither can action explanations be adequately replaced by any collection of causal explanations.

Yet, we can rely on action theory to bring more inferential power to interventionist models, by treating rationalizations that unify stages of an action *as if* they were causal connections. It is important to emphasize that they are not *in fact* causal in character; the normativity of rationalizations cannot be adequately represented in interventionist modeling. Because rationalizations unify in a stronger way than mere causation, such a treatment *under*utilizes rationalization in terms of the inferences that could be justified on its basis. We can treat them as if they were causal, use these connections for making causal inferences, and thereby generate models that can make more predictions about what will happen in such systems.

I argue for this by laying out some key pieces of conceptual machinery that are required for using the approach to causal modeling variously referred to as interventionism, causal Bayes nets, or causal structural equation modeling. The Causal Markov and Causal Faithfulness assumptions are substantive, in that they make non-negligible claims about the underlying nature of the systems being modeled and can be empirically checked to ensure that they are warranted. By committing to these assumptions, we gain powerful techniques for inferring causal structure from probabilistic relationships among variables, and for predicting probabilistic relationships from causal structure. Similarly, the rationalization that explains an action by situating it as a means to another action, constitutes a form of constraint on the causal options available to genuinely rational agents. This constraint can be formalized into an assumption, the Rationalization condition,[1] that can be made about a system of causal variables, in a manner analogous to the Causal Faithfulness and Causal Markov conditions. Thus, Thompson's characterization of the internal unity of action can be incorporated into causal modeling once specific conditions are met.

Section 8.2 lays out a brief overview of naive action explanation and the relation of rationalization that holds between an action performed as the means and the action the performance of which is the end, highlighting the features that will turn out to be useful in incorporating rationalization into causal modeling. Section 8.3 contrasts causal explanation with distinctively mathematical explanation in order to draw a distinction between two ways of applying the model. It is a key part of the trajectory of the overall argument to show that naive action explanation behaves like the modally stronger distinctively mathematical explanations, because of the way it is 'applied' as a model, rather than with the comparatively weaker strength of causal explanation. Section 8.4 introduces the role of conditions like Causal Markov and Faithfulness. Section 8.5 introduces Rationalization as a new condition for causal

modeling. Section 8.6 illustrates the use of the Rationalization condition with the example of driving. Section 8.7 concludes.

## 8.2 Naive action explanation and rationalization

This section lays out a brief overview of Thompson's naive action explanation, examining the character of rationalization as a relation that situates the action to be explained as a means towards or stage in another overarching or more encompassing action.

Thompson begins by identifying a characteristic pattern of explanation involved in actions. Following his lead, I will use the example of baking bread. Suppose someone walks into the kitchen, sees you reaching up into the cupboard, and asks why. We often explain the action of reaching up into the cupboard by situating it as a stage, means, or part of a more encompassing action, like getting the flour. Getting the flour is itself a means or stage that can be explained with recourse to the more encompassing action of making bread. Naive action explanation thus explains by situating the explanandum as an action that is a smaller part of a larger structure that subsumes it and the other requisite action-stages *as* stages of the larger action. One is doing A as part of doing B; one is doing B, then C, then D, as part of doing X. There is a nested structure: getting the flour is itself comprised of smaller actions, like reaching up, grasping, pulling, carrying, and so on. But getting the flour is then a means to starting the dough, and starting the dough is itself given further naive action explanation as a means to the end of baking bread.

Such explanation relies on the 'in order to' that connects the more concrete and limited action to the goal or overarching action into which it fits as a stage. In baking bread, the overall action is not one can *do* except by doing other actions. One bakes bread *by* getting the flour, adding the ingredients, kneading, letting it rise, and so on. There is no separate action of baking the bread that is additional to or separate from the instrumentally performed actions of kneading, rising, baking, and so on (a well-known point since Ryle (1949)).

The relationship that bears the explanatory load in naive action explanation is that of rationalization. An action like getting the flour has a special relation to the action of baking bread. It is not merely that both actions happen to be going on, nor is it that engaging in the first action causes one to engage in the other; it is rather that the first is done specifically because it is a stage in the second. The performance of the first action is in service to the performance of the second. It is only because of this relationship that explanatory illumination can be shed on the first action by situating it with respect to the second. This cannot be a straightforwardly causal relation: starting the dough by no means causes one to later knead the dough, or allow it to rise.

In explanation via rationalization, both relata are actions. They could not be otherwise, in order for it to be a relation of rationalization, rather than some other kind of relation. An explanation that involved an action as a relatum, either as explanans or explanandum, but involved merely a causally defined second relatum could not possibly be a naive action explanation. This is not to say such explanations cannot exist. It is to say that they would not qualify, by the very nature of naive action explanation, as an example of such explanations. Rationalization as an explanatory relation can only hold between two actions.

Rationalizations, on Thompson's view, can be given a non-final form: one action can be performed in service of another, without that further action being somehow a final end or an overarching and self-complete end in itself. Thus, we can find that action A might serve as a stage in the unfurling of a larger action B, which is itself just a stage in some further action C. B can rationalize A, in providing a naive action explanation of it, without thereby having to ground that in some final action. Action B may rationalize A; B may in its turn be rationalized by C (see Chapter 5, Section 5.2, in particular). B provides explanatory traction on A even though it may be incomplete considered as an explanation required to capture everything about action A. B need not be some final or end action, some not-itself-naively-explained action, to provide substantive explanatory work with respect to A. Rationalization of a means by an end action can explain without the end itself having to have some special quality of finality, or to be further judged in terms of its legitimacy to be undertaken. Even if we don't think someone should be baking bread right now, it is nevertheless the fact that they are baking bread that provides the explanation of their reaching for the flour.

This has the consequence of blocking calls for complete finality in allowable ends. The rationalization of kneading the dough as a means to the end of baking bread does not need to culminate with yet further naive explanation of how baking the bread then fits into some action of being healthy, or enjoying a hobby, or living a fulfilled life, and so forth. The end of having baked bread already rationalizes the stages, without further termination. We can simply explain one action by another, if it fits in the right way, and thereby have improved on our explanatory situation, even though the explanans action clearly itself could be a further explanandum. This feature will allow it to fit neatly into causal modeling, as we see in subsequent sections.

Naive action explanation cannot simply be a new type of causal explanation. There is nothing in starting bread dough that *causes* one to subsequently let the dough rise or bake it. Yet knowing that someone has started bread dough does license one to infer that they will be letting it rise and baking it later on. In such a case, it is not the rationalizing action of baking bread that is the direct subject of the inference. I might infer you are baking bread by noting that you are kneading dough, using

naive action explanation; but it is not the same kind of relation that obtains when I note that you are kneading dough and infer that in an hour or two you will be baking it. Baking the dough is also a stage or means towards the end of baking bread, along with kneading the dough. This highlights how one can infer to future actions that are means of the same action: that two actions are rationalized by the same end action provides an inferential handle that connects them as means of the same end. This inferential connection between two actions that are means rationalized by the same action will, in Section 8.5, provide the foundation for using rationalization in causal modeling.

## 8.3 The model versus the system as primary target of inquiry: Comparing distinctively mathematical explanations and naive action explanation

With this account in hand, this section turns to contrast rationalization and naive action explanation with causal connection and causal explanation. By the end of this section, I aim to have shown that action explanation is deployed in a manner closely analogous to distinctively mathematical explanations rather than causal explanations, in terms of how models and systems fit together. This in turn means that rationalization in naive action explanation offers a modally stronger degree of connection than does mere causal explanation.

Lange (2012) defends the claim that there are certain kinds of explanations, which he calls distinctively mathematical explanations, that have a distinctive degree of necessity and cannot be assimilated to causal explanation without loss. One example is that of a mother with 23 strawberries and 3 children. There is no way to evenly divide the strawberries among the children without cutting the fruit. The mother's failure to divide the strawberries evenly among the kids is, however, not merely some causal fact: it is not that she lacks a knife, or is counting incorrectly, or otherwise causally prevented from doing so. Lange points out that it is the mathematical fact that 23 is not evenly divisible by 3 that does the explanatory work. Even though it is something about the physical world being explained, rather than a purely mathematical fact, it is a mathematical explanation and not a physical one involving causation.

Andersen (2018) responds to Lange's claims in several ways. The key response that I want to redeploy here is to make a distinction between two ways in which a model can be used. These reflect two different kinds of modeling tasks, with different orientations towards fitting a model to a system (Andersen 2018). In brief, one way to use a model for a system is such that the system being modeled has priority in determining what is 'wrong' when there is a failure of model-system fit; and in the second kind of modeling tasks, the model itself has priority as an object of study, such that a system which fails to fit the model is rejected in

search of systems that do fit the model. These are both legitimate modeling tasks – it is not that one should be endorsed over the other. Rather, it highlights how taking a different primary focus in terms of the object of study – the system being modeled or the model being used – leads to two different kinds of explanations of the system in question from the model in question. First, I will illustrate this in a scientific case, and subsequently, apply it to naive action explanation.

Consider the Lotka-Volterra toy model. The Lotka-Volterra (LV) equations give the population of a predator and prey population over time. The population size of either at a given time is a direct consequence of the birth rate and death rate at a previous time increment. For the prey population, the death rate is a function of the predator population at the relevant time. For the predator population, the birth rate at a later time is a function of the earlier prey population. This model is a very useful example of a toy model: it is known as being a very simplified, idealized, and often numerically inaccurate model of actual predator and prey populations. Much of the failure to be numerically accurate stems from the fact that very few systems actually fit the model – it is hard to find genuinely isolated predator and prey populations that meet the conditions for these equations to fully apply. Nevertheless, they are extremely useful.

Sometimes, such well-developed toy models can be studied on their own, since many different scientists, with very different target systems, might use versions of it. The equations treated as a toy model can be used to derive the robust Volterra principle (Weisberg & Reisman, 2008). This states that when a general biocide event (something that kills both predator and prey indiscriminately) occurs, then in the recovery period afterwards, the proportion of prey to predators goes steeply up. This turns out to be a mathematical result of the model: any simultaneous increase in the death rates can be shown to result in this change in proportion. It falls out as a purely mathematical consequence of the equations. It is useful and interesting to know of the LV equations that they have this feature, even if it turns out that no actual system ever follows those equations strictly.

This illustrates the distinction between two ways of applying a model: taking either the target system or the model itself as the primary focus of inquiry. In the first way of applying a model to a system, a particular system is being modeled. If the assumptions do not fit that system, the model must be rejected. The system comes first, and the model must be tailored to fit that system. Many cases of modeling are like this. The wildlife biologists in charge of managing some specified conservation area will often have just this kind of focus. The ecosystem(s) are fixed, in that they are well specified as the target requiring a model for the purposes of, e.g., prediction of future population changes. If there is a general biocide of some kind and this change in proportion of prey to predators is not

observed, one goes looking for another model other than LV. It doesn't disprove that the general result holds for LV; it demonstrates that the LV model does not fit the system.

In the second way of applying a model, the model itself is a focus for inquiry. The LV equations can themselves be studied, as clearly illustrated by the way in which Weisberg and Reisman derive the robust Volterra principle. In this kind of modeling task, one starts with the model and goes looking for a suitable system that it fits. It turns out that a case of chemicals dumped in the sea near Italy illustrates this general biocide result effectively; the ratio of prey fish to sharks shot up in the recovery period. If, however, the example from Italy ended up not fitting the model, then we could simply move on to look for some other system that better illustrates the effect. We would not, in this approach, reject the LV as not applying and continue modeling the chemical dump system. We would look for a better fit by taking the model with us and leaving that particular system behind.

What is extremely important in this contrast between model usages is that we already knew that the robust Volterra principle would obtain in any system of which the model held, before we ever even found a system of which the model held. It *had to* hold of any system of which the equations hold, because it is a straightforward mathematical consequence of the equations of that very model. This does not guarantee that we would ever find such a system of which the model holds. But it does ensure, with mathematical certainty, that *if* we find a system of which this model holds, *then* that system must also obey the robust Volterra principle. It holds with mathematical certainty, and nothing weaker, for the systems of which it does end up holding.

Causal explanations are generated when a model is applied in the first way. When we focus on the system in question first, the LV equations help us track the causal relationships governing changes in one population with respect to the influence from the other population. Causal explanations have some degree of strength of connection; they are not merely accidentally true generalizations, for instance. But since they are empirically dis/confirmable like this, they do not hold with mathematical necessity; mathematical necessity is stronger than causal connection.

Distinctively mathematical explanations are generated from the model applied in the second way. Metaphorically, it is like we are walking around with a bag, into which we only put a certain kind of stone. We know that there will only be that kind of stone inside the bag, because we ensured that it would be so by using it as a selection criterion. We don't need to check each stone already in there to make sure the contents of the bag fit the criterion; we *enforced* the criterion in the first place. It might turn out that the bag is empty, because we have not come across any such stones yet. But we know with certainty that *if* there are ever any stones in the bag, they will be of that kind, because we will only put that

kind in. In the second approach to modeling, we enforce the criterion that the system must fit the model that is the focus of inquiry, such that it must be the case that all systems that turn up success are already known to have certain features.

All of this is set up to make the following point: in action explanation, especially in naive action explanation, naive action explanation is treated akin to the LV model applied in the second way. We can usefully explore the features of action by treating it like a model that is a target of inquiry. Then, we go looking for examples that fit the 'model' of action by rejecting those that don't and looking until we find examples that do fit. If we discover that a particular example turns out to not be an action, for whatever reason, we have two options, mirroring the two approaches to modeling. We can distinguish psychological explanation as taking the first kind of approach, where we reject action explanation as providing sufficient traction on the example, but stick with the example and resort to merely psychological explanation instead of action explanation. Or, we take the second approach by sticking with action explanation and rejecting that candidate as not an action, and continue the search for some better example that is an action. Naive action explanation, by dint of holding between two actions, must pre-select for action; it cannot, by definition, end up holding of non-actions. This enforced pre-selection criterion ensures that anything that can be said of action explanation will hold, in systems of which it holds, with a strength like that of mathematical explanation, and not like a mere causal explanation.

Action explanation enforces the selection criteria, like enforcing the criterion of only putting stones of a certain kind in the bag. As a consequence of this, it must be the case that whatever ends up in the action bag is already known to have certain features, which can be explored by taking action itself – in this case, naive action explanation – as the target of inquiry. The existence of behaviors that are not actions, are neither here nor there for that purpose; it merely means that we pass by those examples as we engage in naive action explanation. Thus, we can know things about any case of genuine action that we find in the world prior to ever finding it, by dint of the fact that we can draw inferences from the 'model' itself, studying action. This means that rationalization, as the relation that unifies actions performed as means to an action that is also an end, will yield explanations that are modally stronger than merely causal explanation.

## 8.4 How causal Markov and faithfulness justify inferences in causal modeling

We turn now to see what makes the engine of interventionist or Causal Bayes Nets modeling work. These techniques are essentially a set of algorithms to make justified inferences between probabilistic relationships

in data and causal structure as represented in structural equations and directed acyclic graphs (DAGs). The inferences work with the assistance of some background assumptions or conditions that provide the justificatory foundation for those inferences. In cases where these assumptions fail to hold, we would be unjustified in making inferences that require them. When those assumptions fail, only limited versions of the algorithms can be used, resulting in weaker available inferences; the strongest inferences can be made when the full set of conditions hold. When modeling a system, we ought (in the epistemic sense of ought) to use the strongest set of assumptions we are justified in believing obtains for the system in question.

One central assumption in causal modeling is the Causal Markov condition. This condition stipulates that causes are probabilistically independent of their non-effects, conditional on their parents (Hitchcock, 2018; Pearl, 2009; Spirtes, Glymour, & Scheines, 2000; Woodward, 2005). Put another way, after conditioning on the parents of a given variable, then the only remaining probabilistic dependencies are effects of that variable. This condition allows us to make the inferences that are fundamental for causal modeling: using intervention to distinguish causal from correlational structure. Without conditionalizing on the parents of a target variable, any other effect of those same parents will be probabilistically correlated with the target variable, even though they are not an effect of it. By conditionalizing on the parents of a cause, the dependencies with non-effects are 'broken' for common case structures. In a nutshell, then, the Causal Markov condition ensures that existing conditional dependencies are due to causal relationship(s) and not coincidence.

What would it look like if the Causal Markov Condition failed? How empirically substantive is this condition? This has been the focus of some back and forth (Cartwright, 1999; 2002; Hausman & Woodward, 1999; 2004). Part of what emerged from this disagreement is that Causal Markov is a genuine assumption about the world. It could fail, if we found that there were persistent probabilistic dependencies between variables that could not be accounted for by causal connections. The correlations would have to be both robust over time, and genuinely inexplicable with respect to causal connection. It would be a pure 'spooky' correlation. Cartwright emphasized that this assumption is not trivial, a priori, or merely analytic in character. Hausman and Woodward emphasized that this is an assumption most of us are willing to commit to without much by way of metaphysical misgivings. An upshot is that the Causal Markov condition is empirical, in that we can genuinely consider what it would be like to find that it is violated somewhere, but also metaphysical, in that considering how it would be violated requires rejection of the principle of sufficient reason, for instance.

Another central assumption, the Causal Faithfulness condition, is also a key part of licensing inferences between causal structure and

probabilistic relationships in data. It ensures that existing conditional probabilistic independencies reveal causally independent variables. Faithfulness is a feature that a directed acyclic graph (DAG) may have to a set of probability relationships among the variables in that graph. A graph is faithful to the probability distribution when there are no causally connected variables in the graph that are independent in the distribution. Put another way, the causal faithfulness condition ensures that there are no 'hidden' causal dependencies that fail to show up in the probabilities in the data.

The Causal Faithfulness condition can be violated when there are precisely counterbalanced causal relationships that 'disappear' by being probabilistically independent in the data despite being causally connected in the true graph. Consider a cause C that has two pathways by which it is connected to effect E, one path that brings about C with a weight of. 8, and another path where C causes D with weight 1 and D then suppresses C, with weight -. 8. C and E are causally connected in the true graph, and if the weights of those pathways were anything other than precisely opposite, they would be probabilistically dependent in the data. But because the two pathways have exactly opposing weights, so that C causes E with precisely the same strength that it suppresses E, it looks as if C has no influence on E.

If the parameter values for causal relationships in graphs were randomly distributed, then this violation would occur with measure 0 frequency (Spirtes et al., 2000). But this is itself a substantive assumption about systems. Zhang and Spirtes (2008) argue that it may be violated in mechanisms like thermostats for maintaining a fixed room temperature. Andersen (2013) argues that it may be violated even more rampantly, since evolved systems use homeostatic mechanisms that are much more finely tuned than thermostats, and which are evolved precisely to maintain homeostasis. Modified versions of the condition, which would fail less often to hold of systems though also support somewhat weaker inferences, can be used (Forster, Raskutti, Stern, & Weinberger, 2017; Zhang & Spirtes, 2016).

Taken together, the Causal Markov and Causal Faithfulness conditions ensure that the probabilistic dependencies and independences in data taken from a given system connect in reliable ways with the causal relationships in the true causal graph of that system. Without these, one cannot infer between data and *causal* relationships. Even though Causal Faithfulness and Causal Markov assumptions require substantive commitments about the systems in question, they return advantages in making genuine discoveries. Thus, making the strongest set of assumptions about Faithfulness and Markov that are warranted by the particular system being modeled allows us to use the strongest version of the inferential tools that are justified by those conditions.

## 8.5 Introducing the rationalization condition

The outcome of this section will be the introduction of a new condition, an addition to the two most commonly used in current practice. This assumption, the Rationalization condition, will ordinarily be violated: for the overwhelming majority of systems, it will fail to hold, and the default modeling apparatus is used. But there do exist systems in which the Rationalization condition holds. And in modeling such systems, use of this condition will strengthen the inferences we can make, in particular from the causal graph to predict probabilistic relationships in data. Insofar as we should use the strongest available set of inferences given the conditions that we are justified too by the conditions that obtain in the system(s) being modeled, this Rationalization condition is a useful way to add power to interventionist causal modeling.

My proposal, put very briefly, is that we add what I will call the Rationalization condition: variables representing distinct actions that are each rationalized as means of the same end have that shared rationalization treated as a causal arrow between them, where the causal order follows the temporal order of those actions. The rationalizing explanation, the naive action explanans, is not directly represented in the system with a variable. Only the two naive action explananda are represented. They are connected via the Rationalization relation with an arrow in the graph if and only if they are rationalized by the same (missing) action.

By treating rationalization relations that obtain between appropriately defined variables *as if* they are causal in the context of causal modeling, we can predict additional probabilistic relationships in the data. Since, as we saw in the previous section, rationalization is a stronger explanatory connection than causation, it can be weakened to mere causation without thereby overextending our justificatory base. Refraining from using the Rationality condition simply reverts to the same causal modeling techniques we currently use.

Begin with actions that can be naively explained by a further action. Distinct means to the same end may be used to define variables such that those two means variables are treated as if causally connected. There is a two-dimensional figure, that of the rationalization relation situating the first action as a means in an end action, and the second action as means to the same end action. This is projected onto a one-dimensional arrow connecting the two means actions. The two means actions must be sufficiently distinct that one can occur without the other thereby occurring; the temporally earlier means action is treated as cause and the temporally later means action as effect. In this regard, the causal graph must simplify rationalization in a way that loses genuine structure. This both illustrates why causal relations could not be used to in general reduce rationalization relations, but also how such projection provides the requisite inferential basis to justify inferences about connections between those two action variables.

Key to using Rationalization is that two different stages of a single action can be explained with respect to that same action: if we ask why I am kneading dough, and ask why I am letting it rise, both are naively explained with respect to the same baking of the same bread. They are each distinct stages of that same single action, temporally differentiable means to the end of making bread.

In a system of variables that includes ones like Kneading Bread and Letting Dough Rise, these variables will be probabilistically connected: it is an empirical fact about the world that when we identify genuine instances of each of these two variables, using that action description, they will be consistently positively correlated. Yet it is also clear that they are not straightforwardly *causally* connected, in the way that dropping a glass and getting the floor wet are causally connected. Instead, these variables are connected via rationalization: an instance of each variable is given a naive action explanation rationalizing each with the same action. Kneading bread does not count as *causing* one to let it rise. Even the weaker sense of causal connection is lacking – nothing compels that connection, even weakly, except the aims of the agent performing them. But they go together with such consistency that it can be reliably used in prediction: when someone is Kneading Dough, it is quite likely that later they will let it rise. It is rationalization as uniting these as two means to a common end that provides the connection and the prediction, not causation.

Thus, we can add an arrow in the directed acyclic graph (DAG) between these two variables, just as if it were a regular causal relation, and makes inferences in terms of predicting probabilistic relationships in the data taken from such a system. We can thus predict, and explain, the systematic correlations we find between these two variables, by incorporating this additional arrow in the graph. Refraining from using the Rationalization condition in such cases won't lead to a model that makes inaccurate predictions. But it will lead to a model that generates weaker predictions and explanations than it could. We would be refraining from saying true things that we could say if we relied on the Rationalization condition.

What it takes to make this rationalization condition part of the formal apparatus is straightforward. The action variables should be defined so as to allow for fairly straightforward identification of instances of the variable. This is a generic feature of the craft aspect of modeling, and not particular to action. For these variables, a superscript R is added, indicating that a given variable is being used with respect to its rationalization relationships. It need not *only* be used for rationalization connections, but it *may* be so used. This superscript is then also added to the weight of the connection in the DAG. Thus, some variables, and some causal arrows in a DAG, will have an appended R superscript to remind us explicitly of the requirements for their use in the model.

Even when using the Rationalization condition, the independence of causal variables (e.g., Campbell, 2010) must be ensured. If there is a

single occurrence in the world, for any kind of system, which ends up counting as an instance of two different variables in a system of variables, then those variables will appear causally connected even to those they are not. To avoid double-counting, variables are defined to ensure that no single instance counts as an instance of both of them. This is implemented in systems involving Rationalization by leaving out any variable representing the shared end action, and only including the common means to that end as variables. Kneading Bread and Letting Dough Rise will be independent variables in the appropriate way. Kneading Bread and Making Bread will not: some or most instances of Kneading Bread will also be instances of Making Bread. But for different actions that are stages or means in the same action for a naive action explanation, the conditions of variable independence will be met.[2]

The Rationalization condition will not be met in the vast majority of causal systems being modeled. When we think this condition is violated – when, for any number of reasons, we lack a genuine agent that could potentially offer genuine naive explanations of their actions – we simply don't treat the relations between stages of an action as causally linked. Humans will often fail to be the kinds of systems where we can assume that the Rationalization relation holds. Just as there are systems where Causal Faithfulness fails, this situation simply means that the additional analytical tools based on that assumption cannot be deployed. In modeling human behavior that fails to be adequately rational, our predictions made using the Rationalization condition will be less accurate than in systems where it holds. But this does not indicate that the Rationalization condition is fundamentally unusable. If one is modeling predator-prey relationships, then the Lotka-Volterra model might be apt. If one is not modeling such relationships, there is no reason to even consider the L-V model; yet surely the fact that there are modeling situations where it doesn't apply does not mean that the L-V model is *never* correct. Just as surely, that there are many cases where humans fail to be agents in the requisite way does not mean they are *never* agents in the requisite way.

By only relying on the Rationalization condition when we are justified in holding that genuine rational action is taking place within the system to be modeled, we are using a selection method for systems. By enforcing the appropriate selection criterion for systems to which we apply the rationalization condition, we know before we ever find such a system that certain features will obtain and can draw on this in making inferences.

## 8.6 Using the rationalization condition: Turning left

There are many, many instances of actions that turn out to be quite prosaic but which clearly demonstrate that not only *could* we use naive action explanation in prediction much like using causation, but that we

already *do* this, so effectively that it provides a cornerstone of modern living: driving.

Consider first the general structure of driving somewhere in the context of naive action explanation. Imagine we are driving down a particular street, and someone asks, "Why are you driving down this street?" Our responses, in cases like this, are not of the form that a causal explanation would require, even a very general or abstractly described one. It would be weird and tedious to answer such a question by saying that we had been driving on the street back there, and turned right at the corner onto this street, and that was why we were driving on this one. Knowing the path that we took to be on this very street just does not answer the question of *why* we are on it; instead, it answers something more like *how* we got to it. More fitting answers involve situating our driving down this street in the larger encompassing trajectory of our drive. We are driving *to* that place over there, and this is the only connecting street between where we were and where we are going. We are driving to some further destination and thought this was a more scenic road than the other alternatives. Our map directed us here as part of the fastest journey from starting point to destination. And so on.

All of these are kinds of naive action explanation. We take our driving down the street to be akin to reaching for the flour on the shelf, and explain this part, driving on this very street, by encompassing it into a trajectory that presupposes our end in driving is to arrive at a set destination. We *are here* because we are going *to there*. Any particular part of the drive is explained as a stage in a longer drive defined by end points. We consider it to be the exceptional case when we really aren't going to anywhere, just driving around for no reason. It is only in such cases that there is no unifying action under which to subsume our current driving. Indeed, we still usually give such explanations a naive flavor, explaining this drive with respect to the absence of such an arrival-at-destination plan into which it is a stage, situating it into something like an entertainment-or-diversion end instead.

Consider next how turn signals, used properly, display a driver's intentions so that we consistently rely on them to predict how to safely navigate roads shared with other drivers. If I am at a stop sign, and a car on the other side is also stopped, and we both have our left turn signals on, I confidently pull into the intersection when there is space, on the knowledge that the other car will not be driving straight (and thus, into my own car) but instead turning left. I don't have to see inside the window to the driver, much less peer into their secret intentions, in order to predict what they will do next. At this particular stage of their drive, they intend to turn left, and we confidently, and with high success, predict that they will be turning left when the opportunity such as a break in oncoming traffic arises. Recall the non-finality of actions that may serve

as rationalizing ends, from Section 8.2. We need not know where they are *really* going to know that they are turning left here.

When facing, at a stop sign or red light, a car with a left turn signal blinking, nothing about having a turn signal on *causes* the driver to turn left. Yet having the turn signal on, and subsequently actually turning left, are both highly correlated. Interestingly though unsurprisingly, having no turn signal on is less highly correlated with going straight than having a left or right signal on is correlated with actually turning left or right. Driving behavior demonstrates very clearly that we already have a myriad of ways of thinking about genuinely intentional behavior, replete with actions and full-fledged intentions, about which we have no qualms reasoning.

If one wants to model a system for car movements in a given intersection, one could include variables like Turn Signal [Left, Right, None], Car Movement [At rest, Continues straight, Turns Left, Turns Right], Stoplight [green, yellow, red], etc. There will be R superscripts on the first two variables, but not on the Stoplight variable. There will be a causal connection between Car Movement and Stoplight, which need not be treated with a superscript R. The stoplight is not an R variable: a requirement for using the Rationalization relation is that it holds only between two R labeled variables. Between Turn Signal and Car Movement, there will be a new arrow added to the graph, with an R superscript on it. Adding the superscript R allows for a few additional variables and arrows to be introduced to a graph that would not otherwise be possible, and which allow for more predictions about behavior in the system to be made.

Recall from the section on naive action explanation (Section 8.2) that such rationalization relations can only transpire between two actions. This limits the extent to which such additional arrows will be added to the DAG. They can only connect R variables, of which there will be a limited number as well. If each and every variable in a system is an R variable, then in a real sense one is not doing causal modeling, and should switch to using straightforward naive action explanation instead. It is only when mixing clearly causal variables with action related variables that the Rationalization condition will come in handy.

In general, we must treat other drivers as if their driving-related actions are causally connected, in order to rely on each other to follow traffic rules and thus stay safe on the road. We already know, when pressed, that these are not strictly causal relations in the way that a dropped glass of water and a wet floor are causal. But we have to interact with other drivers at a mass scale that makes it easier to treat these *as if* they were causal. A breakdown in road rules, where we cannot predict what other drivers will do, leads to a worse situation for everyone; that is what happens when R fails.

What happens in modeling such cases when the Rationalization condition fails? Two comparisons are illuminating. First, compare this to

failures of the Causal Markov and Faithfulness conditions. If we have inadequate justification to treat the system as containing actions subject to naive action explanation, then the Rationalization condition either just isn't used, or holds trivially by not applying to any of the variables or arrows. If there is only one R variable, then there will be no R relations in the graph, and no need to invoke Rationalization.

Second, compare the failure to the two ways of applying a model, from Section 8.3. If we suspect that we have a case where there are genuinely no actions susceptible to naive action explanation (perhaps we are looking at badly programmed driverless cars), then we have two available moves. We could take the first approach to using a model and reject Rationalization, sticking with the system at hand and developing some other set of variables to better reflect its causal structure. Or, we could take the second approach, and reject the system: if we want to model genuine driving behavior, we would reject such a system and continue on to find a better system that illustrates the Rationalization condition.

## 8.7 Conclusion

The internal connection between means and end exhibited in naive action explanation has a modal strength that is more like that of distinctively mathematical explanations than that of causal explanations. Yet, because it can be treated in DAGs, and meets criteria like D-separation, it can be used to strengthen inferences that can be drawn from causal models. This chapter aimed to motivate incorporation of the Rationalization condition into causal modeling practices, where it is apt for the system(s) being modeled, and to provide the basics for incorporating R variables into systems of variables and R arrows into DAGs.

The proposal developed here fits in a longer trajectory of discussion of action and causation that goes back to Davidson (1963) and Anscombe (1981). Since Kim's (1998) Causal Exclusion problem, the issue of causation and action, or causation and the mental in any guise, has been construed in terms of causation relating higher and lower levels, rather than competing descriptions of the very same relata, as Davidson originally discussed. It has also led to a widespread sense, mostly among those working in the philosophy of science, that interventionist causal modeling has supplanted any genuinely causal role for something as ephemeral and internal as reasons or action. The ways in which many philosophers have attempted to eliminate or reduce action explanation using interventionist or related causal modeling approaches involves a deep misunderstanding of the character of action.

Once we note that rationalization and causation behave differently, we could decide to reduce action and insist it be replaced with causal explanation. The fact that kneading dough does not cause one to let it rise could mean that there is nothing more to connect them than the

tenuous possibility of some weak physical causal chain. On the other hand, we could use modus tollens instead of modus ponens and conclude that the failure of causation to accommodate the connection between stages of actions like bread-making means that causation itself is insufficient to handle such a connection, and look to *supplement* causal analysis with action analysis where it is apt. Reliance on the Rationalization condition where it is appropriate can be justified by its own usefulness. It also paves a better path forward in bringing together these distinctive forms of explanation to enhance rather than replace one another.

## Acknowledgements

Much thanks for helpful discussion and comments from Shimin Zhao, Varsha Pai, Matthew Maxwell, Cem Erkli, Zili Dong, and Weixin Cai. Thanks to Michael Brent for editorial improvements. I am grateful for the opportunity to live and work on the unceded territory of the Musqueam, Squamish, Tsleil-Waututh, and Kwikwetlem nations.

## Notes

1. It is key to distinguish between beliefs and action: the rationality of actions, given in their rationalization relations as they unfurl from beginning towards completion, is the specific target here, not an epistemological notion of rationality that applies primarily to beliefs.
2. This proposal thus differs from other extensions of causal modeling, such as Schaffer (2016): instances of grounding will violate the independence condition, and fail D-separation, in a way that is avoided by the Rationalization condition.

## References

Andersen, H. (2018). Complements, not competitors: Causal and mathematical explanations. *The British Journal for the Philosophy of Science*, *69*(2), 485–508.

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, *80*(5), 672–683.

Anscombe, G. E. M. (1981). *Metaphysics and the philosophy of mind*. Minneapolis, MN: University of Minnesota Press.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, *20*(1), 64–79.

Cartwright, N. (1999). Causal diversity and the markov condition. *Synthese*, *121*(1), 3–27.

Cartwright, N. (2002). Against modularity, the causal markov condition, and any link between the two: Comments on hausman and woodward. *The British Journal for the Philosophy of Science*, *53*(3), 411–453.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.

Forster, M., Raskutti, G., Stern, R., & Weinberger, N. (2017). The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, *69*(3), 821–848.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance, and the causal markov condition. *The British Journal for the Philosophy of Science*, *50*(4), 521–583.

Hausman, D. M., & Woodward, J. (2004). Modularity and the causal markov condition: A restatement. *The British Journal for the Philosophy of Science*, *55*(1), 147–161.

Hitchcock, C. (2018). Probabilistic causation. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/.

Lange, M. (2012). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*, *64*(3), 485–511.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.

Ryle, G. (2009/1949). *The concept of mind*. London: Routledge.

Pearl, J. (2009). *Causality*. Cambridge, UK: Cambridge University Press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies*, *173*(1), 49–100.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT Press.

Thompson, M. (2008). *Life and action*. Cambridge, MA: Harvard University Press.

von Wright, G. H. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.

Weisberg, M., & Reisman, K. (2008). The robust volterra principle. *Philosophy of Science*, *75*(1), 106–131.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. New York, NY: Oxford University Press.

Zhang, J., & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, *18*(2), 239–271.

Zhang, J., & Spirtes, P. (2016). The three faces of faithfulness. *Synthese*, *193*(4), 1011–1027.

# 9 Skepticism about Self-Understanding

*Matthew Boyle*

> History consists of the thoughts and actions of minds, which are not only intelligible but intelligent, intelligible to themselves, not merely to something other than themselves.
>
> R. G. Collingwood, *The Idea of History* (1946, p. 112)

## 9.1 Introduction

Although much in our lives is opaque to us, it is nevertheless true that a central strand of our thought and action has a certain seeming intelligibility. Suppose you and I have an argument about whether a certain philosophical theory is sound. I vigorously maintain that it is, but your objections linger with me, and gradually I become persuaded that you are right. I become convinced that the theory won't do, and I seem to know why: I know what considerations changed my mind. Or, to take a more mundane example, suppose you ask me whether I want to come along to the beach. I am at a phase in life where this involves daunting preparations: wrestling uncooperative children into swimsuits, applying oily sunscreen while they struggle to escape, finding lost shoes, etc. The prospect of all this is almost too much to contemplate. But then I think of the strange bravery that comes over my children when they finally are in the water, the yell of triumph my daughter lets out when she is hit by a wave, and I decide to come along after all. Again, I seem to know why I decided as I did, what persuaded me. I could talk about it at some length, if anyone were interested.

In these and similar ways, we normally stand ready to answer questions about why we believe what we believe and do what we do.[1] It is true that we sometimes find our own thoughts and choices mystifying. Still, we are, to a striking degree, prepared to offer, without self-observation or inference from other evidence, explanations of our own beliefs and choices. We appear to presume that we normally know our own grounds for belief and action, and indeed, that we do not merely know them, but are – in a sense that is not easy to clarify – in charge of their influence. This is why we count these judgments and decisions as,

DOI: 10.4324/9780429022579-10

in Collingwood's elegant phrase, "not only intelligible but intelligent": they do not just strike us as comprehensible; we take them to proceed *from* our conscious acceptance of the adequacy of certain reasons. I do not merely suppose that I *know* what caused me to hold a given philosophical view, or what caused me to come along to the beach; I suppose that I myself was *responsible* for the action of the relevant causes, which would not have operated without my assent. We could express this point by saying that our normal understanding of our own judgments and decisions purports to be *directive* rather than *post hoc*: it purports to be the kind of understanding possessed by someone who knowingly *makes* something so for certain reasons, rather than the kind of understanding possessed by someone who merely *takes* something to be so for certain reasons, without supposing his so taking things has an influence on the matter.

My interest in the philosophical topic of mental action grows out of an interest in our capacity for such comprehending self-determination – for "making up our minds", as it is often put. Whether "mental action" is really a helpful frame through which to approach this topic is open to question: it is not clear that such self-determination is primarily a matter of doing things *mentally*, rather than simply a matter of leading our tangible, non-mental lives in distinctively self-governed ways. But at any rate, I think a significant part of recent philosophical interest in mental action has grown out of a concern with our capacity for such self-governance, construed as a capacity actively to shape our own judgments and decisions.[2]

In this essay, I want to focus, not on some positive theory of such self-governance, but on an increasingly influential form of skepticism about whether we really possess the relevant sort of self-understanding at all. The skepticism I have in mind draws on well-known results in social and behavioral psychology. Psychologists have devised a variety of ways of showing that our professed reasons for judging and choosing are often not our true reasons, and that we are ready to offer confabulated rationales with no more self-examination, and no less assurance, than we exhibit in more straightforward cases. These observations have led a number of psychologists and philosophers of mind to suggest that our sense of immediate self-understanding is an illusion, not just in certain cases, but in general. According to Hilary Kornblith, for instance,

> What appears in introspection to be the direct apprehension of causal relations among our mental states is really, at bottom, the result of a process of rational reconstruction: we are actually engaged in a subconscious process of theorizing about what the source of our beliefs must have been.
>
> (Kornblith, 2013, pp. 198–199)[3]

Skeptics about self-understanding like Kornblith maintain that, in many cases, we are mistaken about what explains our own beliefs and actions. But even when our explanations are correct, they argue, the process by which we arrive at our account is essentially the same as the one operative in cases of confabulation: we *reconstruct* our grounds, on the basis of general assumptions about what the causes of our beliefs and actions are likely to be. This activity of self-interpretation may not occur consciously: it may be performed subconsciously by a dedicated "mind-reading faculty." But however evident our explanations may seem, they are at bottom interpretations resting on a kind of speculation about the causal determinants of our thoughts and actions, rather than on any immediate awareness of these determinants. In this way, such skeptics call into question, not merely the reliability of our self-understanding, but whether it can ever be what it purports to be: an immediate apprehension of our own guiding reasons for belief and action.

This skeptical attitude toward human self-understanding has a growing number of advocates in academic philosophy, but I think it is not just an academic phenomenon. The unreliability of our conscious self-understanding has become a popular theme of newspaper op-eds and magazine think-pieces, and has been the basis for a number of "pop psychology" best-sellers.[4] What is noteworthy about these books is not just what they say, but that they are so popular: it suggests that we – at any rate, a growing number of us – have an appetite for this sort of thought about ourselves. We are ready to be skeptical of our own pretensions to self-understanding. We feel an obscure satisfaction in seeing the mask of Reason torn away.

I will not speculate about the sources of this satisfaction: this strikes me as an interesting question, but not one I possess the tools to answer. My aim in this essay will be simply to consider the extent to which our naïve assumption of self-understanding is undermined by the kinds of experimental results cited by skeptics, and what can be said on the assumption's behalf. I will argue the grounds for skepticism are overstated, and that the naïve view has a deeper basis than its critics recognize.

## 9.2 The case for skepticism

*Skeptics about self-understanding*,[5] as I will use the term, are philosophers and psychologists who hold that we do not have any immediate awareness of our own grounds for belief and action. The case for such skepticism often begins with the claim that we are frequently mistaken about our own grounds for belief and action, but the skeptic's ultimate target is not just a claim about how often we are correct, but a certain intuitive conception of how these matters are known to us – namely, that they are somehow immediately or transparently available. I will have more to say later about how we might understand such transparency.

Skeptics generally do not spend much time trying to characterize it, since they doubt that there really is any such mode of availability. Their focus is on arguing that we have no reason to believe in such transparency, whatever it might amount to.

To make matters concrete, I will focus on the version of skepticism presented by Peter Carruthers in his *The Opacity of Mind* (2011), which gives an unusually detailed and sophisticated presentation of the case for skepticism. Carruthers offers many kinds of evidence for his position, but I will focus on one main strand in his argumentation, a strand that also figures prominently in the work of other skeptics.[6] The points I make in response could, I believe, be adapted, *mutatis mutandis*, to apply to other common skeptical arguments.

The argument on which I want to focus begins from the observation that, in various kinds of circumstances, people can be induced to *confabulate* about their grounds for belief or action – that is, to produce, with apparent assurance and sincerity, demonstrably false accounts of their own reasons for belief and action. Thus – to begin with an extreme case – "split brain" patients who have undergone a cerebral commissurotomy, in which the connection between the two hemispheres of the brain is severed by bisecting the corpus callosum, can be brought to offer confabulated accounts of their own actions. Carruthers describes a case in which a commissurotomy patient was shown a card with the word "Walk!" on it, presented in such a way that the card could be seen only by his left eye (which transmits signals exclusively to the right brain hemisphere). The subject then stood up and began to leave the testing space, and when asked where he was going (a verbal inquiry processed by the speech centers in his left brain hemisphere), replied, "I'm going to the house to get a Coke." In this and many similar cases, it is natural to conclude that what such subjects offer as their grounds for action are not their true motives, but *post hoc* rationalizations invented in response to a demand for an account where the true account is unavailable. What is striking about such cases, however, is that the confabulating subjects are in many instances conscious neither of any uncertainty about their motives nor of any effort of self-interpretation. Their awareness of their grounds for action seems to them just as direct as it is in ordinary, unproblematic cases.

What leads commissurotomy patients to confabulate is, of course, a rare and pathological brain condition, but Carruthers argues that such confabulation can also be induced in subjects whose brains are perfectly normal. This seems to be the lesson of vast body of experimental evidence accumulated over the last several decades by social psychologists. The experiments take a variety of forms, but their general structure can be stated as follows: the experimenters produce a

situation in which the judgments or choices of experimental subjects are significantly affected by some surprising factor, and then interrogate the subjects about why they judged or chose as they did. The subjects prove to be strikingly unaware of the role that the surprising factor played in affecting their choice – indeed, in many cases, they are incredulous at the suggestion that this factor played a role. Instead, they cite, with apparent assurance, factors that cannot plausibly have played a decisive role. Hence their accounts of why they judged or chose as they did are regarded as instances of confabulation.

As an illustration, Carruthers mentions a widely-cited experiment described by Richard Nisbett and Timothy Wilson in their influential paper "Telling more than we can know" (1977). In a mall in suburban Michigan, Nisbett and Wilson conducted what purported to be a consumer survey, in which they asked shoppers to inspect four identical pairs of nylon stockings arrayed on a table, and to determine which was "the best quality." After the subjects had made a choice, they were asked why they chose as they did. A pronounced "position effect" was observed, such that the farther right stockings were in the array, the more frequently they were chosen as of highest quality, and the right-most pair was preferred to the left-most pair by a factor of nearly four to one. This was, in fact, predicted by the experimenters, since "position effects" on choice constitute a well-established phenomenon in the psychology of decision. Yet when the subjects were asked why they chose as they did, none mentioned position, and when asked whether position had played a role in their judgment, nearly all of them confidently denied it (with the exception of one subject who was taking a psychology course that had recently discussed position effects). Instead, they pointed to attributes of the preferred pair, such as its "superior knit, sheerness, or elasticity" (Wilson, 2002, p. 103). But given that the stockings were in fact identical, it is not credible that such attributes played a decisive role. Nisbett and Wilson take this and numerous similar results to show that "the accuracy of subject reports about higher order mental processes may be very low" since "such introspective access as may exist is not sufficient to produce accurate reports about the role of critical stimuli in response to questions asked a few minutes or seconds after the stimuli have been processed and the response produced" (Nisbett & Wilson, 1977, p. 246). And since the proffered reasons were not in fact decisive factors, such results are also commonly taken to show that we readily confabulate about our reasons for judgment and choice (cf. Wilson, 2002, Ch. 5; Carruthers, 2011, Ch. 11).

Carruthers draws two conclusions from these sorts of results. First, he concludes that our subjective impression of having immediate awareness of our own grounds for judgment and decision must be taken with a grain of salt. As he puts it:

> [W]e don't have any subjectively accessible warrant for believing that we ever have transparent access to our own attitudes. This is because patients can report plainly-confabulated explanations with all of the same sense of obvious and immediacy as normal people.
>
> (Carruthers, 2011, p. 43)[7]

Secondly, he takes the results to show that, in general, people have the ability rapidly and subconsciously to construct interpretations of their own behavior. The confabulated explanations offered by brain-damaged subjects clearly draw on such an ability, and the kinds of results described by Nisbett and Wilson seem to show that subjects whose brains are normal can also be induced to construct confabulated self-interpretations in quite ordinary sorts of circumstances.[8]

Once we have admitted that we have such a self-interpretative faculty, however, it is natural to ask whether it might be operative, not just in cases where confabulation is conspicuous, but also in cases where everything seems normal. Carruthers calls this the "universal mind-reading hypothesis" (UMRH). In general, according to UMRH, we acquire beliefs about our own attitudes and their causes by subconsciously framing interpretations of the available data – data which may include facts about our history and circumstances, observation of our own words and deeds, and awareness of our own "phenomenally conscious" states. Sometimes our self-interpretations are inaccurate, but even when they are accurate, the explanations we offer do not reflect any immediate insight into our own reasons for judgment and action. We have no such insight: the only advantage we have over others, in interpreting ourselves, is access to some data to which they are not privy.

UMRH is not ruled out by our intuitive sense that our awareness of our grounds for judgment and decision is sometimes immediate, for subjects who undoubtedly confabulate – such as split-brain patients and people duped by social psychology experiments – also characteristically take themselves to have immediate awareness of their grounds. Moreover, UMRH provides an attractively simple and unified explanation of how we arrive at an understanding of our own beliefs and choices. The only alternative, seemingly, is to accept a "dual method" hypothesis (DMH), according to which we *sometimes* form views about our own reasons for belief and action by relying on a self-interpretative faculty, but on other occasions, we become aware of them in some more immediate way.[9] But once our subjective impression of having immediate knowledge of our own reasons has been discounted, there seems to be no strong reason to prefer DMH and a great deal of evidence in favor of UMRH. Thus, Carruthers concludes, UMRH is probably true, and so our sense of immediate self-understanding is probably an illusion.

## 9.3 Processualism about self-understanding

Carruthers's case for skepticism might be questioned at various points, but here I want to focus on a general background assumption that he and other skeptics make about the nature of self-understanding.

The assumption that interests me is expressed in the opening lines of Nisbett and Wilson's "Telling more than we can know", a paper that is a touchstone for skeptics:

> "Why do you like him?" "How did you solve this problem?" "Why did you take that job?" In our daily lives we answer many such questions about the cognitive processes underlying our choices, evaluations, judgments, and behavior.
>
> (Nisbett & Wilson, 1977, p. 231)

The questions with which Nisbett and Wilson begin evoke a familiar kind of situation: one in which people are asked about what *persuades* them to do something, think something, or feel a certain way – that is, why this way of acting, thinking, or feeling seems to them reasonable, or at any rate seems to have some reason that speaks in favor of it.[10] Nisbett and Wilson immediately go on, however, to characterize these questions as concerned with "the cognitive processes underlying our choices, evaluations, judgments, and behavior", and they set about assembling evidence that we are often mistaken about the factors that influence these processes. That is, they represent subjects who answer these questions as claiming insight into what *precipitated* or *brought about* a given thought, attitude, or action. They assume that, when we answer questions such as "Why do you like him?" or "Why did you take the job?", we are offering an account of some *process* – some connected sequence of events unfolding over time that led to the advent of the relevant attitude or act. I will refer to this assumption as *processualism* about self-understanding.

Processualism can seem like an inevitable consequence of the thought – which I do not dispute – that self-understanding is a form of causal understanding, an understanding in which we trace judgments and decisions to the factors that explain their existence. It is this seeming inevitability, presumably, that accounts for Nisbett and Wilson's unargued transition from speaking of answers to explanatory questions to speaking of cognitive processes. I want to suggest, however, that this is a substantive and dubious transition, and that once we query it, the case for skepticism about self-understanding loses much of its force. For only if processualism is true does evidence that we are often ignorant of important factors in the causal history of our attitudes support skepticism about self-understanding.

Suppose you have a somewhat abrasive friend, and some other friend asks, "Why do you like him?" Perhaps you might answer that you know he can be abrasive, but you think this just reflects his discomfort in social situations, and over the years he has proved to be loyal and kind when it counts. The details here don't matter; the thing to notice is simply that, to the extent that it is right to represent your answer as citing psychological causes of your present affection for your friend, these causes are not past but present mental states. You like him because you *think* his abrasiveness is superficial and you *take* him to have been loyal and kind when it counts. It is your *present* conviction on these points, and your *presently* taking them to speak in favor of your affection, that you put forward as explaining your liking him. Of course, the present you describe is not an "instantaneous present": you aim to characterize a standing relationship between an attitude you hold and other things you take to be true, a complex cognitive state which has existed for some time and persists into the present. In this sense, your explanation may have implications, not just for the present moment, but for your reasons for *having liked* this friend for some time. Your explanation does not, however, commit you to any claims about the cognitive *processes* by which your affection arose: it describes what (presently) *sustains* your affection, not the causal processes that *brought it about*.

One feature of such explanations that may lead us to overlook their non-processive character is the fact that they often cite facts about the past as grounds for present attitudes. The foregoing explanation, for instance, cites the fact that your friend *has been* loyal and kind when it counts. But here it is important to distinguish between the rational *ground* of an attitude and its proximate psychological *cause*. What makes it reasonable to feel affection for this person is a fact about the past: that on earlier occasions, your friend showed himself to be loyal and kind. But this fact forms the content of a *present* mental state: you *take* your friend to have been loyal and kind in the past. And it is your presently taking this to be so, not the sheer past fact, that explains your attitude. It is true that ordinary explanations of attitudes often leave such present mental states implicit: they simply cite facts about the past, allowing the subject's present awareness of these facts to be implied by general presuppositions of explanatory relevance. But it should be clear, on reflection, that a fact about the past that rationally supports a certain attitude can figure in a reason-giving explanation of the attitude's presently being held only if this fact is presently known to the attitude-holder.

To bring out the importance of this requirement of present awareness, it will help to consider a case in which the requirement is not met. Suppose it is true that

(E) You like NN because he has been loyal and kind to you

but the relevance of his loyalty and kindness to your present affection is not accessible to you when you consider the question of why you like him. What sort of explanation of your affection could (E) be in that case? It could not be an explanation of what you now find *likeable* about him, what *speaks in favor* of liking him from your perspective, for what you do not know cannot persuade you of something. (E) must, then, be a different kind of explanation, one that asserts your present affection for NN to *result* from certain experiences you had in the past. But if this is what (E) means, it is surely not the kind of explanation we normally presume ourselves to be able to give, without self-observation or inference from other evidence, of our present attitudes. For there is no reason to presume that I will remember psychological factors that figure in the causal history of my attitudes; and even if I do remember them, the claim that their occurrence is causally relevant to my present attitude is hardly one on which I would be able to comment without further investigation, were it not for the fact that they operate through my regarding them as reasons *to* feel affection for my friend. But if these experiences affect my attitude through my presently taking them to support my attitude, then they operate by being the content of contemporaneous awareness, and this awareness explains my attitude non-processively.

It might be objected that this analysis cannot apply to other questions on Nisbett and Wilson's list, such as "Why did you take that job?" and "How did you solve this problem?" These questions invite the addressee to explain an event that is itself in the past: how can this be explained by a present appreciation of a case for holding an attitude? – Well, it is true that these questions invite the subject to explain, not why she is presently persuaded to do something or how she thinks something can be done, but why she was persuaded or how she did it. But although the fact that the *explanandum* is in the past, this does not imply that the *explanation* must appeal to a process. When I explain why I took a certain job, I appeal, presumably, to my memory of what *persuaded* me to take the job. But though this decision is now in the past, it was once made in the present, and I would then have been able to explain it by citing things I *presently* believed, wanted, thought, etc. My ability to speak, now, about my reasons for past decisions, to the extent that I have one, surely depends on my ability to remember how the world looked to me then: I am able to speak for my reasons because I am able to project myself back into my earlier point of view, the one from which I made the decision. But to the extent that I can do this, I can explain the decision as if I were making it now – by speaking, from the point of view I remember, to the question of what grounds I see *for* making the decision.

Something broadly similar applies in the case of my ability to answer the question "How did you solve this problem?" If the question is understood simply as a request for an account of the sequence of thoughts, images, etc. that led up to my discovery of the solution, then there is little

reason indeed to expect that I will remember everything relevant, or that I will be competent to judge what caused what. But there is another reading of the question, one that naturally occurs to us when it is listed together with "Why do you like him?", "Why did you take that job?", etc. On this other reading, the question is understood as a request to explain what I grasped about how *to* solve the problem. My answer to this question may take the form of a narrative ("Well, I saw that if you multiply through by the denominator then you eliminate the fraction, and then I noticed …"), but what I am really recalling is what I understood about how the problem can be solved, and narrating this understanding in a sequence of steps. This is not primarily an account of the *process by which I came to discover* my solution, but of my *solution* itself.

These considerations suggest a principled reason why our ability to answer questions about the grounds for our own attitudes must appeal to present mental states and events. It is only in virtue of the fact that I can treat the question why I hold a given attitude as "transparent" to the question whether *to* hold this attitude that I am in a position to answer it without self-observation or inference. But the latter question is not a question about whatever events may have brought about my attitude, considered as a psychological reality, but about the *case* for holding the attitude, considered as a possible response to some question of attitude-transcendent fact. This case can indeed be transposed into a psychological register: instead of answering "Why do you like him?" by saying "Well, *p* and *q*", I might equally answer "Well, I know that *p*, and I think that *q*." And I see no reason to deny that, if true, this psychological rendering of the explanation gives a causal account of my affection for my friend, by citing other attitudes that sustain it. But if my appreciation of a case for holding a certain attitude is to explain my holding the attitude, the relevant explanation must travel through my capacity to appreciate the reason-giving force of a case for an attitude, and this requires that I grasp the case as a whole and appreciate its import at one time.[11] Thus the transparency of first person attitude explanations implies their non-processuality.

## 9.4 Processualism and skepticism

I will say more about such transparency in Section 9.5. In the meantime, let us consider the bearing of these observations on skepticism about self-understanding.

Skeptics commonly take evidence that people are ignorant of factors that have played an important role in bringing about their judgments and decisions as evidence that they lack insight into why they make the relevant judgments and decisions. In one sense, this is surely correct: the experimental results, if valid[12], show that factors to which we are

oblivious may influence our judgments and decisions; and it is plausible that such factors operate, not just in artificial experimental conditions, but in ordinary situations in which we judge and decide. That factors such as the ordering of objects in an array can influence our preferences among them is disconcerting enough, and it is all the more unnerving to know that we are ready to offer quite unrelated rationales for such preferences. But do such observations show that the explanations we then offer are false and confabulated?

The skeptical conclusion that they do must rest, it seems to me, on something like the following line of thought:

> *The fact that a certain pair of socks was on the right in the presented array was crucial to producing the subjects' preference for this pair. But the explanations subjects offer of their preference makes no mention of its position; and furthermore, their belief that the socks have the attributes appealed to in their explanation (superior knit, sheerness, elasticity, etc.) can itself be plausibly explained only by a pre-existing disposition to prefer socks on the right-hand side of the array. So the explanations these subjects offer must not be the true explanations of their preferences; they must rather be confabulated.*

Our discussion of processualism equips us to notice an equivocation in this line of thought. There is compelling evidence that subjects may be unaware of a factor that plays a decisive role in *bringing about* their preference, and it is implausible that the factors they do cite are the real determinants of this process. Does it follow that they do not know why they have the relevant preference? Only if the relevant "why?"-question is understood as an inquiry into the process by which their preference came to exist; for only then does their ignorance of what was decisive into this process imply that they are ignorant of what explains their preference. But in fact, as we have seen, the relevant sort of "why?"-question inquires into a different topic: what considerations *sustain* the subject's attitude in the present. This is not an inquiry into the causal history of her attitude, but into the broader outlook that supports her holding it.[13]

For all these experimental results show, the belief that (e.g.) this pair of socks is of superior knit might indeed have been the basis on which a given subject prefers it – though it might also be true that the pair's being on the right-hand side of the array brought it about that he had this belief (perhaps by leading him to attend to its knitting with a more favorable eye). So such experimental results do not prove that people do not know the "real reasons" for their preferences, though they may show that our perceptions and judgments can be influenced by factors of which we are ignorant and whose influence we would not countenance if we were aware of it. But our ignorance of the factors that played a decisive role in the *process* by which our preferences arose does not show

that we are ignorant of what *persuades* us to prefer this pair, for this is not a question about a mental process. To draw this distinction is not, of course, to offer any positive proof that the accounts such experimental subjects give of their preferences do express what really persuaded them; but as far as I can see, the kinds of experiments standardly cited by skeptics fail to show that they do not.

This may seem too quick. Recall the case of the split-brain subject who was shown a card that said "Walk!" and explained his getting up and walking by saying that he was "going to the house to get a Coke." Surely his decision to get up and walk (or his voluntarily beginning to walk, if there was no decision) preceded his inventing this rationale. It seems obvious, here, that the rationale was concocted under pressure to explain already extant facts, and was thus a textbook case of confabulation. But if it was a confabulated account of extant facts, how can it be any kind of genuine explanation, processive or otherwise, of those facts? If the decision to walk pre-existed any thought of going to the house to get a Coke, then this thought can't explain the decision – not even in the sense of being what non-processively persuaded the subject to walk. And if we grant this point, it seems that a similar argument will apply to the preferences in the Nisbett and Wilson experiment: the subjects find themselves preferring the socks on the right side of the array, and under pressure to find a reason for their preference, they seek a feature that would justify it. That they come to believe the socks possess this feature – superior knit, or whatever – surely reflects an already-existing tendency to find socks on the right-hand side preferable. But then their judgment that the socks have this feature can't be a genuine explanation of their finding these socks preferable, for their preference pre-existed any such judgment.

I think this objection still depends on processualist assumptions. Let us grant for the sake of argument that the split-brain subject begins to walk before thinking of going to get a Coke and that the subject who prefers socks on the right-hand side does so before becoming convinced of their superior knit. On these assumptions, it is certainly true that the thought of going to get a Coke does not *bring about* the subject's walking, and that the judgment that the knit of these socks is superior does not *bring about* the subject's preference for them. But again, this is not the sort of explanation we are seeking when we ask a subject why he *holds* a certain preference or *is performing* a certain action. We are asking for an explanation, not of what brought about a judgment or choice, but of what *sustains* it here and now – why the subject presently holds the relevant view or is willing to carry on with the relevant action. For all these experiments show, it might be that the pressure of a "why?"-question from the experimenter brings about a crystallization in the subject's outlook, such that she is *now* walking to the house with the aim of getting a Coke, or *now* prefers these socks on the basis of their knit. Whether

the preference or choice *pre-existed* what presently guides or sustains it is neither here nor there.

This might seem like a pyrrhic victory over the skeptic.[14] If we deny that rationalizing explanations make claims about the processes by which judgments and decisions come to exist, are we not insulating such explanations from criticism by depriving them of any real significance? Are we not, indeed, conceding the skeptic's point in all but name? Skeptics maintain that our judgments and choices are often influenced by factors of which we are unaware and that the rationales we offer for them are *post hoc* rationalizations rather than accounts of the bases of which we *consciously make* the relevant judgments and choices. If we reply that these rationalizations may nevertheless come to *sustain* the relevant judgments and choices, isn't this just more grist for the skeptic's mill? Doesn't it amount to conceding that our capacity to reflect on our reasons for judgment and choice operates, not as a power consciously to determine what we judge and choose, but merely as a factor that tends to rationalize and reinforce determinations which arise from other, non-conscious factors?

Again, I think this retort seems compelling only if we are still in the grip of processualist assumptions. Let me make two points in response. First, the fact that rationalizing explanations speak to what sustains judgments and choices, rather than to what brought them about, does not trivialize these explanations. To represent an attitude as non-processively explained by other attitudes is to assert that a subject's being in one *state* depends on her being in other *states* she is in, and this mode of understanding can only get a grip where there is a certain stable relation of dependence between aspects of the subject's outlook. There is a real and significant form of explanatory dependence here, one that may persist *through* time, though it does not consist in a causal process unfolding *over* time. To assess whether such a relation holds, we need to consider, not how the relevant attitudes originated, but how the continued existence of one is related to the continued existence of the other. Does the subject who says she is "going to the house to get a Coke" actually follow through on this project? Is the judgment of the subject who says she prefers these socks because of their knit responsive to changes in her assessment of their knit? It is characteristic of these experimental situations that nothing of significance hangs on the relevant attitudes and projects, so their robustness in counterfactual conditions is likely to be weak. But in principle, these sorts of counterfactuals are what must be tested to assess a claim about what *sustains* a preference or *guides* an action.

Secondly, the impression that results like those reported by Nisbett and Wilson show that our rationales for our own judgments and choices must in general be *post hoc* rationalizations, rather than expressions of directive awareness of what determines our attitudes, presupposes what

we might call a "ballistic" picture of the causation of judgment and choice. In the normal course of things – though not, typically, in the kinds of situations produced in these experiments – we make judgments and choices about topics that *matter* to us, topics concerning which our being wrong or unreasonable has a cost. Moreover, we must *persist* in these attitudes through some period of time – the time it takes to carry out a complex project, or the period during which we persist in holding the view expressed by a certain judgment. The *sustainability* of our views and projects requires that our overall view of things – of what is true and what is worthwhile – has a certain order and coherence.[15] Only if we thought that particular judgments, preferences, etc., were carried forward by their own momentum, without need for support from the rest of our outlook on the world – only if our picture of the explanation of judgments and choices were in this sense ballistic – could we suppose that our views about why the relevant judgments and choices are sound are mere rationalizations of no real explanatory significance. But the kinds of experimental results we have been considering establish no such thing.

The kinds of results reported by Nisbett and Wilson are certainly unsettling: they show how easily our judgments and choices can be influenced by factors of which we are unaware. But I do not see how they could show that our judgments and choices are in general independent of our (relatively stable, relatively coherent) overall view of what is true and what is important. If they are thus dependent, however, then any given judgment or decision, whatever its origin, will normally be able to persist only insofar as it can find a place in our general view of what is (knowably) the case and what is (intelligibly) worthwhile. In this way, our conscious reasons for accepting given judgments and choices *matter* to the existence of these judgments and choices. Indeed, when seen from this perspective, our tendency to "rationalize" our own judgments and choices takes on a different aspect: it appears as the operation of our power to consider the relation between given judgments and choices and the background of attitudes that sustains them. The fact that this power can, in certain kinds of cases, be induced to rationalize attitudes for which there is no pre-existing rationale does not show that, in general, it is irrelevant to our holding the attitudes in question.

## 9.5 Transparency and self-understanding

So far, I have simply argued that the kinds of observations cited by skeptics fail to show that our naïve assumption of self-understanding is false. But what can be said in support of this assumption? I'll conclude with some remarks about this.

Earlier we noted a close connection between our ability to offer explanations of our own attitudes and our ability to treat the question of why

we hold a given attitude as "transparent" to the question of whether *to* hold such an attitude – a question whose focus is, not the *explanation* of an extant attitude, but what *speaks in favor of holding* a given attitude toward some topic. For instance, in answering the explanatory question of why I reject a certain philosophical theory, I can normally treat this as equivalent to the justificatory question of why the theory is to be rejected, and convert my answer to the latter question into an answer to the former by transposing the justifying reasons I would offer into claims about the beliefs that contribute to explaining my judgment. And similarly, if I am asked to explain why I chose to come along to the beach, I can normally treat this as equivalent to the justificatory question of what made it choiceworthy to come and transpose my answer into an explanatory register by rephrasing my rationale for coming as an account of the background of attitudes that explains my choice to come.[16]

This kind of transparency of an explanatory question about an attitude to a justificatory question about how to view the world (henceforth, "explanatory transparency") is not trivial, for the two questions are distinct on their face: the one concerns the causes of my own psychological states; the other, the right or appropriate attitude toward some non-psychological topic.[17] Moreover, explanatory transparency does not always hold: it is a familiar aspect of the frailty of human rationality that we sometimes cannot bring ourselves to hold an attitude that we judge to be appropriate, or cannot bring ourselves to give up one that we recognize to be unsupported. But to the extent that I cannot see a justification for a given attitude, I will also not take myself to be able to explain my own attitude in the characteristically immediate, non-speculative way on which skeptics seek to cast doubt. So the case where explanatory transparency holds is the one that should interest us: why does this relationship hold, when it does?

It seems to me that the relationship is grounded in the nature of rational intentional attitudes as such. In general, we can think of intentional acts and attitudes – judgments and beliefs, choices and intentions, preferences, desires, hopes, fears, etc. – as stances on some characteristic question.[18] Thus a belief that *p* is an affirmative stance on the question whether *p*; an intention to $\phi$ is an affirmative stance on the question whether to $\phi$; a preference for A over B is an affirmative stance on the question whether A is preferable to B; a desire for O is an affirmative stance on the question whether O is desirable; a hope that *p* is an affirmative stance on the question whether *p* is to be hoped for; a fear of X is an affirmative stance on the question whether X is to be feared; and so on. If we define a rational animal as an animal that can *think* about such questions – one that can consider them in an interrogative mode, and deliberate about what speaks for a given answer to them, rather than simply coming unthinkingly to accept some answer – then we may say that, for rational animals, holding an attitude will, in general, involve

being *disposed* to think of the relevant question in a certain way.[19] This will be true even if they do not actually deliberate about the corresponding question: even if they never consider this question as such, their attitude will involve some standing conception of why the question is rightly so answered, which will connect their answer to this question with other things they hold true, desirable, choiceworthy, etc.

If the attitudes of rational animals are states of this sort, then such animals will in general have a special kind of awareness of their own attitudes: an awareness "from the inside", so to speak. For their holding a given attitude will itself involve their being disposed to answer some corresponding world-directed question in a certain way on a certain basis. This, I believe, is what accounts for the phenomenon of explanatory transparency. What accounts for such transparency is not that we "make up our minds" at the moment when a question about our own attitude arises, and know the grounds on which we have reached this new assessment.[20] This may occur in certain cases, but the phenomenon of explanatory transparency is more general: in many cases in which we already hold a settled attitude on some topic, including cases in which we have never reflected on the attitude in question, we take ourselves to be able to explain why we hold the attitude by speaking to the question what speaks in favor of holding the attitude in question, i.e., to the case for holding it. What accounts for this phenomenon in its full generality, I suggest, is that, for rational animals like ourselves, holding an attitude just *is* being disposed to find a certain answer to a question compelling on a certain basis (or, in the limiting case, primitively). There are many different ways of finding an answer to a question compelling, of course, and in many instances what makes a given answer seem compelling will be only a very vague conception of how this answer fits into our overall understanding of the world and our own bases for making determinations about it. But this is no objection to our account of explanatory transparency: it represents our ability to explain our own attitudes as no stronger, but also no weaker, than it in fact tends to be.

If this account of explanatory transparency is sound, however, there can be no basis for a *general* skepticism about the self-understanding of rational animals. It may be that we are easily misled about the explanation of our own attitudes and that our attitudes can be influenced by factors of which we are unaware and whose effects we would not endorse. Still, insofar as we are capable of holding intentional attitudes at all, this will in general involve our finding answers to corresponding questions compelling on certain grounds (or primitively), and this stance will enable us, on reflection, to offer real explanations of the relevant attitudes, by converting the case we see for a given answer into an account of the background that sustains our attitude. A subject who offers this sort of explanation will be speaking *from the perspective* of her own attitude, articulating the wider worldview that makes a certain answer to

an attitude-defining question seem compelling. What she will then offer is indeed an explanation of her attitude, but it is an explanation that is integral to the first-order outlook she describes, rather than a mere hypothesis about certain extrinsic causes of her attitude.

It follows that there can be no basis for a general skepticism about such self-understanding, since the relevant understanding simply makes explicit commitments involved in the subject's holding the relevant first-order attitudes themselves. Consider again our subject whose affection for her abrasive friend rests on his record of loyalty and kindness at moments that matter. On our analysis, her attitude toward her friend *consists* in her taking him to be likeable – worthy of affection – on these grounds. Hence, when she makes these grounds explicit in an explanation of her affection, she will not merely be offering a hypothesis about what brought her affection into being, but characterizing the particular way of answering the question of his likeability in which her attitude consists: there may be countless other ways of finding a person likeable, but this is hers. The explanation she offers of her affection will thus characterize her holding of this affection intrinsically, rather than merely identifying certain independent causes of her holding this attitude. Of course, any number of confounding factors may lead her to mischaracterize the explanation of her own attitude, but such mischaracterizations, however frequent they may be in practice, are secondary in principle. For insofar she holds the relevant attitude at all, the primary explanation of her holding it will necessarily be available to her, provided that she is not distracted, deluded, or otherwise misled.

## 9.6 Conclusion: Intelligibility and intelligence

It is not a new idea that human beings, as rational animals, are distinguished from nonrational animals by the fact that we lead our lives with a certain implicit self-understanding. One classic expression of this thought is the claim that the "historical" or "human" sciences (*Geisteswissenchaften*) are set apart from the natural sciences (*Naturwissenschaften*) by the fact that they seek a different kind of understanding of phenomena, not lawlike explanation (*erklären*) but "interpretative understanding" (*verstehen*). The nature and defensibility of any such distinction is, of course, controversial, but a recurring theme in the writings of Dilthey, Collingwood, and other defenders of the distinction is that humanistic understanding seeks to capture the *internal standpoint of a participant* on the events, institutions, and practices under consideration. Thus Collingwood writes (commenting approvingly on a theme he finds in Hegel):

> [H]istory consists of actions, and actions have an inside and an outside; on the outside they are mere events, related in space and

> time but not otherwise; on the inside they are thoughts, bound to each other by logical connexions… [T]he historian must first work empirically by studying documents and other evidence; it is only in this way that he can establish what the facts are. But he must then look at the facts from the inside, and tell us what they look like from that point of view.[21]

Collingwood's thought is that the sort of understanding sought by the historical sciences is one that takes account of the *self-understanding* of the parties involved. He holds that an adequate understanding of human affairs must comprehend this "internal" perspective because he assumes that human events are shaped, in the main and on the whole, by our own self-understanding. This is what he means when he says, in the passage quoted as an epigraph to this paper, that human thoughts and actions are "not merely intelligible, but intelligent": they are not merely comprehensible by others *post hoc*, but guided *ab initio* by the subject's own understanding. Hence any adequate interpretation of these phenomena must take account of this perspective.

Our discussion enables us to see a valuable point in this way of thinking. The point is not merely that human beings characteristically have views about their own grounds for thinking and acting, but that, in the fundamental case, these views merely bring to reflective articulacy a standpoint that constitutively informs the relevant thoughts and actions themselves: a standpoint whose primary focus is, not the explanation of our own thoughts and actions, but questions about the world in which we think and act. Such understanding is directive, rather than merely *post hoc*, inasmuch as it is a condition of the existence of the relevant attitudes and actions: not an extrinsic, productive cause of their coming into being, but an intrinsic, sustaining cause of the stance in which they consist. It is the fact that our attitudes and actions consist in ways of answering such questions that ensures that human beings have a perspective "from the inside" on their own thoughts and actions, and that this perspective is, not just a well-informed piece of speculation, but a standpoint comprehension of which is necessary for understanding the relevant thoughts and actions themselves.

This certainly does not imply, and Collingwood did not take it to imply, that the understanding people have of their own thoughts and actions must always be taken at face value, or must exhaust what there is to understand about human events. It implies only that, for any such thought or action, there will be some standpoint on a question which belongs to it constitutively, and which can, in favorable conditions, be transformed into an explicit self-understanding. Comprehending this standpoint is not the end but at most the beginning of understanding why people think and act as they do. Nevertheless, it is a crucial

beginning, inasmuch as it is essential to a full understanding of why the relevant thought or action exists and what it means.

This might seem an overly rationalistic conception of human thought and action. Isn't it true that we often think and do things for which we can offer no justification, or indeed which we take to be unjustified? Of course we do, but this is not in tension with the conception of self-understanding defended here. I have admitted that the limiting case of self-understanding – a case which may in fact be common – is one in which we have no particular reason for an attitude or action, or none beyond our general sense that it seems primitively true or attractive. Even when we admit we have no particular reason for a given attitude or action, we lay claim to self-understanding, inasmuch as we claim to know what our own reasons are. To claim that one has no specifiable reason for a given attitude or action is to assert that one *does* have a perspective "from the inside" on the relevant attitude or action, and that what this perspective reveals is precisely that the relevant attitude or action is primitively compelling.

I have also not claimed that a person's attitude on a question is always identical to her all-things-considered judgment on that question.[22] A person may, for instance, judge that, all things considered, it is not desirable for him to have another beer – perhaps because he anticipates that he will say things he shouldn't, that he will regret it tomorrow, etc. – yet he may still very much want another beer. But this is perfectly compatible with my claims about the connection between holding an attitude and possessing self-understanding. A person who, despite his better judgment, wants another beer normally *does* have an internal perspective on his desire. He does not find himself blurting out "Give me another!" as if driven by an alien compulsion. On the contrary: he can speak – in whatever minimal way – to what seems desirable about having another beer, and his finding the prospect desirable on this basis just *is* his desiring it. He may not understand why the attractions of this prospect triumph over his considered judgment, but he understands why he desires what he does: he understands the perspective from which he finds having another beer desirable. In this way, even recalcitrant attitudes, which resist our better judgment, can, and normally do, constitutively involve self-understanding.

Perhaps there are also attitudes and actions from which we are alienated in a more radical sense, ones which really do present themselves as alien compulsions whose grounds are inscrutable to us. I do not find it easy to think of examples of this kind from my own life, but this does not convince me that such alienation is impossible. What is impossible, if my argument is sound, is that this sort of alienation should be the rule, rather than the exception, in the life of a rational animal. For a rational animal is one that can think about the questions that are the topics of its attitudes, and such thinking is possible only where, as a rule at least,

the views such thinking expresses are the ones the animal in fact holds. In this sense, although our self-understanding may be fallible and corruptible in all kinds of ways, its foundations are secure insofar as we are as rational, thinking beings at all. And that we are such beings is, as Descartes famously observed, not something any of us can easily place in doubt.[23]

## Acknowledgments

I am grateful to Michael Bishop and Matthew Soteriou for inviting me to the conference from which this paper originated, and to Lucy O'Brien for her illuminating comments on an earlier draft. I am also indebted to Maria Alvarez, Ulrike Heuer, Yair Levy, Richard Moran, David Owens, Sarah Paul, and Sebastian Watzl for astute questions and criticisms.

## Notes

1. In the section that follows, I will consider both our understanding of mental events such as judgment and choice, and also our understanding of mental states such as belief, desire, and intention. Obviously there are important distinctions between these different kinds of mental phenomena, but both raise the question of self-understanding, and the differences between them will not matter to my argument. When I need a generic term for the objects of self-understanding, I will speak sometimes of "thoughts", sometimes of "attitudes", but I intend what I say to apply, *mutatis mutandis*, to both attitudinal states and attitude-expressive events.
2. For examples of this sort of interest in mental agency, see for instance Korsgaard (1996), Burge (1998), Moran (2001), and Hieronymi (2009).
3. Kornblith makes a more detailed case for this outlook in his 2011 and 2018. Other recent authors writing from a broadly similar standpoint include Carruthers (2010; 2011), Cassam (2014), Doris (2015), and Rosenberg (2016).
4. Some instances are: Timothy Wilson, *Strangers to ourselves* (2002); Malcolm Gladwell, *Blink* (2007); Richard H. Thaler and Cass R. Sunstein, *Nudge* (2008); Dan Ariely, *Predictably irrational* (2009); Jonah Lehrer, *How we decide* (2010); Daniel Kahneman, *Thinking fast and slow* (2011); Leonard Mlodinow, *Subliminal* (2013).
5. For brevity, I'll sometimes just call them "skeptics."
6. This strand is the main focus of the briefer case for skepticism in Carruthers (2010).
7. Carruthers's focus is not on our awareness of *why* we think and act as we do, but on our awareness of our own attitudes and choices themselves. But he clearly takes the point to apply also to our subjective impression of having obvious and immediate knowledge of our own grounds for belief and action.
8. Carruthers also holds that this capacity is grounded in the very same "mind-reading faculty" we bring to bear in interpreting the words and actions of other people, but this additional claim won't be crucial here.
9. Cf. Carruthers (2011, p. 45).

10. Looked at from this standpoint, the question "How did you solve this problem?" is an outlier, since (on one reading) it asks the subject to reconstruct the mental process by which she arrived at her solution, which is a question about her actual psychological history rather than about her current reasons. But there is another reading of the question that tends to predominate when it is placed in the context of the other questions: one on which it asks the subject to explain what she has grasped about how *to* solve the problem. And this, again, is a question that invites an explanation of how, according to the subject, the problem can be solved, not a narrative of the thoughts and images that passed through the subject's mind as she considered how to solve it.
11. This is, indeed, implied in the very idea that my *appreciation* of the case explains my attitude: to appreciate a case is not merely to hold a manifold of attitudes toward the several propositions that comprise the case, but to hold some sort of attitude toward the case as a whole, and thus to comprehend the several elements of the case and their import at one time. I say more about this topic in Boyle (2011).
12. There is an ongoing controversy about the replicability of many widely-cited findings in social psychology, but I am not in a position to evaluate this controversy. For the sake of argument, I will take it for granted that the relevant results are valid.
13. Closely related points are made in Malle (2006), though Malle frames his objection in terms of a distinction, which I do not accept, between "reasons" and "causes." Nevertheless, I am indebted to his discussion.
14. This paragraph attempts to capture a concern put to me in conversation by Sebastian Watzl. For a similar objection, see Kornblith (2018, Section 5.3).
15. Not, to be sure, total order and coherence: it is a familiar fact that our judgments may be inconsistent, and that our projects may embody incompatible values. Nevertheless, I assume that a tendency toward order and coherence holds broadly and on the whole, and the idea of a judgment or choice that is wholly isolated from other supporting views about what is true or valuable is at best a limiting case. I say more about these issues in Section 9.6.
16. In some cases, of course, I will have no elaborate rationale to offer for an attitude. For some judgments, I will only be able to say things like "It just seems obvious"; for some choices, I will only be able to say "It just seemed attractive"; etc. But even when I offer only such minimal explanations, I presuppose my ability to speak to the explanation of my attitudes by speaking to the question of why, from my perspective, they appear justified. In this case, I claim that the relevant appearance is primitive. I say more about this point in Section 9.6.
17. In special cases, of course, I may hold an attitude that is *about* some psychological topic, and in this case my answer to the justificatory question will itself concern psychology. The important point, however, is that the justificatory question concerns the *topic* of one of my attitudes, not my *attitude* toward that topic. Having noted this point, I will continue, for simplicity, to refer to the first-order topic as "non-psychological" or "world-directed."
18. This idea is forcefully developed in Hieronymi (2009; 2014). See also my "'Making up our mind' and the activity of reason" (Boyle, 2011).
19. Here and in what follows, I use the phrase "in general" to mark the fact that this principle may admit of exceptions. In cases where what holds in general fails to hold, our holding of an attitude will become detached from

our disposition to think in a certain way about a corresponding question. I do not deny that this is possible, but I deny that it can be the rule in a creature capable of thinking about the questions that are the focus of its attitudes. I say more about this point below in Section 9.6.

20. Moran (2001) is often read – mistakenly, I think – as proposing an account of this sort.
21. Collingwood (1946, p. 118). Cf. Dilthey (1989, pp. 58–59).
22. This paragraph attempts to respond to a question pressed by David Owens.
23. This point bears on a claim made by Alex Rosenberg in a recent *New York Times* Op-Ed summarizing the case for skepticism about self-understanding. Having surveyed the kinds of considerations we noted in Section 9.2, Rosenberg draws the following bold conclusion:

> There is no first-person point of view. Our access to our own thoughts is just as indirect and fallible as our access to the thoughts of other people. We have no privileged access to our own minds. If our thoughts give the real meaning of our actions, our words, our lives, then we can't ever be sure what we say or do, or for that matter, what we think or why we think it.
>
> (Rosenberg, 2016, p. 5)

Rosenberg does not conclude, however, that we have no thoughts, or that our words and actions are meaningless. But if I am right, the question of the existence of these *explananda* and the question of the possibility of the relevant kind of explanation are not separable in the way Rosenberg assumes.

## References


Ariely, D. (2009). *Predictably irrational* (rev. ed.). New York: HarperCollins.

Boyle, M. (2011). Making up your mind' and the activity of reason. *Philosophers' Imprint*, *11*(17), 1–24.

Burge, T. (1998). Reason and the first person. In C. Wright, B. C. Smith, & C. MacDonald (Eds.), *Knowing our own minds* (pp. 243–270). Oxford: Oxford University Press.

Byrne, A. (2018). *Transparency and self-knowledge*. Oxford: Oxford University Press.

Carruthers, P. (2010). Introspection: Divided and partly eliminated. *Philosophy and Phenomenological Research*, *80*(1), 76–111.

Carruthers, P. (2011). *The opacity of mind*. Oxford: Oxford University Press.

Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.

Collingwood, R. G. (1946). *The idea of history*. Oxford: Oxford University Press.

Dilthey, W. (1989). *Introduction to the human sciences* (R.A. Makkreel & F. Rodi, Eds.). Princeton, NJ: Princeton University Press.

Doris, J. M. (2015). *Talking to our selves*. Oxford: Oxford University Press.

Gladwell, M. (2005). *Blink*. New York: Little: Brown and Company.

Hieronymi, P. (2009). Two kinds of agency. In L. O'Brien, & M. Soteriou (Eds.), *Mental actions* (pp. 138–162). Oxford: Oxford University Press.

Hieronymi, P. (2014). Reflection and responsibility. *Philosophy and Public Affairs*, *42*(1), 3–41.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Kornblith, H. (2012). *On reflection*. Oxford: Oxford University Press.

Kornblith, H. (2014). Is there room for armchair theorizing in epistemology. In M. C. Haug (Ed.), *Philosophical methodology: The armchair or the laboratory* (pp. 195–216). London: Routledge.

Kornblith, H. (2018). Philosophy, science, and common sense. In J. de Ridder, R. Peels, & R. van Woudenberg (Eds.), *Scientism: Problems and prospects* (pp. 127–148). New York: Oxford University Press.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

Korsgaard, C. M. (2009). The activity of reason. *Proceedings and Addresses of the American Philosophical Association*, *83*(2), 23–43.

Lehrer, J. (2009). *How we decide*. Boston: Houghton Mifflin Harcourt.

Malle, B. (2006). Of windmills and straw men: Folk assumptions of mind and action. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior?*(pp. 207–231). Cambridge, MA: MIT Press.

Mlodinow, L. (2013). *Subliminal: How your unconscious mind rules your behavior*. New York: Penguin Random House.

Moran, R. (2001). *Authority and estrangement*. Princeton, NJ: Princeton University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. *Psychological Review*, *84*(3), 231–259.

Rosenberg, A. (2016, July 18). Why you don't know your own mind. *The Stone Blog, The New York Times*. Retrieved from https://www.nytimes.com/2016/07/18/opinion/why-you-dont-know-your-own-mind.html

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. London: Yale University Press.

Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: Harvard University Press.

# 10 Embodied Cognition and the Causal Roles of the Mental

*Lisa Miracchi Titus*

## 10.1 Introduction

In this paper I will offer a general defense of the view that the body and environment are metaphysically relevant to cognitive processes: henceforth "Embodied Cognition."[1] Many Classical (i.e., mainstream, non-Embodied Cognition) theorists accept that the body and environment are significantly metaphysically relevant to mental *content*, but maintain that cognitive processes can be fully understood intracranially. I argue here that an adequate understanding of the nature and grounds of cognitive processes must substantially involve the subject's relationship with her body and environment.

Multiple theorists have argued against this thesis on the grounds that the evidence its proponents adduce can be explained equally well by the view that the body and environment are merely *causally* relevant for cognition, not *constitutively*, or *metaphysically* relevant (Adams & Aizawa, 2001; Aizawa, 2007; Block, 2005).

For example, in describing some experimental results Alva Noë (2005) adduces in favor of his enactive view, Ned Block writes:

> These results are impressive but what they show is that sensorimotor contingencies have an *effect* on experience, not that experience is even partially *constituted* by – or supervenes constitutively on – bodily activity.
>
> (Block, 2005, p. 263)

This is an important challenge that I think has not yet been satisfactorily answered to date, and that I will argue gets at the heart of the debate between Embodied and Classical theorists. I offer here a general argument that the body and environment are indeed metaphysically, and not merely causally, relevant to cognition:

1 **Semantic Efficacy**: Mental processes are (generally significantly) inherently content-involving.

DOI: 10.4324/9780429022579-11

2 **Semantic Externalism**: Mental content (generally significantly) metaphysically depends on the body and environment.
3 Therefore, **Embodied Cognition**: mental processes (generally significantly) metaphysically depend on the body and environment.

I use the term "generally" above to allow some specific exceptions, and "significantly" to indicate that the role of the body and environment cannot be ignored or abstracted away in theorizing about how mental content and cognition are metaphysically determined. I use the more general term "metaphysical dependence", even though the term "constitutive" is typically used in framing the challenge. One need not take a stand on the specific kind of metaphysical dependence of the mental to the non-mental in order to adjudicate the debate between Classical and Embodied Cognition theorists, and so I frame the problem more generally here.

I hope this argument will be interesting to the reader for several reasons. First, it is a general defense of Embodied Cognition that shares with Classical Cognition a commitment to realism about mental processes (as opposed to the kind of eliminativism advocated by Chemero (2009), or the kind of instrumentalism advocated by Dennett (1987b)). It also perhaps surprisingly places the focus where Classical Cognition theorists tend to be interested and take themselves to be on firmer ground: on the importance of contentful descriptions in explaining mental states and behavior.

The form of the argument is simple: if mental processes are inherently content-involving, and content is externally determined, then mental processes are externally determined. Inference to the conclusion from the premises relies only on two very plausible commitments. First is the idea that if some feature *F* is a metaphysical determinant of whether *O* has *P*, then where *O*'s having *P* is causally efficacious, *F* is metaphysically relevant to that causal process. Consider what it would be to deny this claim. One would be committed to the possibility of cases where *F* is metaphysically required for *P*'s obtaining but not the obtaining of processes that inherently involve *P*. But processes inherently involving *P* require the existence of *P*, and so feature *F*. So, there cannot be any such case.

The second idea is the transitivity of metaphysical relevance, so that if *A* is metaphysically relevant to *B*, and *B* is metaphysically relevant to *C*, then *A* is metaphysically relevant to *C*. It is hard to find people who disagree with this claim. (I do not have to point out to the reader what a rarity this is in philosophical discussion.) Even for those who deny the flat-out claim – e.g., contrastivists such as Jonathan Schaffer – they accept suitably precisified versions, e.g., that: If *A* rather than *A** is metaphysically relevant to *B* rather than $B^*$, and: If *B* rather than $B^*$ is metaphysically relevant to *C* rather than $C^*$, then: *A* rather than $A^*$ is

metaphysically relevant to *C* rather than *C** (Schaffer, 2012). We could re-frame the argument above to be contrastive without loss of interest or generality. So, very plausibly, the argument is valid.[2]

Semantic Externalism is widely held in philosophy of mind and cognitive science, and I will simply assume it here.[3] Indeed, for this reason, I frame the Embodied Cognition hypothesis as a claim about mental *processes*, rather than states, because it is widely accepted that mental states have their contents essentially and that those contents are largely determined by relations to the body and environment.

The bulk of the work, then, will be in arguing for Semantic Efficacy. While serious work must be done to establish this claim, especially given that its truth in this context entails the minority view of Embodied Cognition, I note at the outset how pervasive and influential this view is in our everyday thought: any time we take ourselves to have *reasons for* believing or doing anything, we take *what* we perceive, feel, think, etc. to be relevant to what we think and do next. I personally do not think I could give up this commitment and still go about my daily life in anything like the way I have been. But people can and do deny Semantic Efficacy for a variety of reasons. It is thus worth defending in detail.

Note that the form of the argument does not depend on particular views about *how* bodily and environmental processes are metaphysically relevant to mental content, or on any particular views about the nature of mental states and processes, beyond the Semantic Efficacy thesis. I will attempt to preserve this generality in my defense of Semantic Efficacy below. This wide scope will usefully clarify and generalize the debate regardless of its ultimate success, given that much of the debate between Classical Cognition theorists and Embodied Cognition theorists to date has to do with details of particular positive Embodied Cognition proposals. I wish to refocus the debate instead on how restrictive the Classical Cognition research program is. In requiring mental processes to be specified intracranially, Classical Cognition distorts our understanding of mental processes and oversimplifies the problem of explaining how they obtain in virtue of non-mental natural kinds. Embodied Cognition, as a general thesis, motivates a more complex and nuanced exploration into these questions.

In Section 10.2, I explain the Semantic Efficacy thesis and discuss why it is actually incompatible with Classical Cognitive science, despite first appearances. This helps to generalize and clarify the debate between Embodied and Classical Cognition theorists. Whereas Classical Cognition claims that mental processes are operations over intracranially specified mental representations, Embodied Cognition holds that mental processes must be specified relationally. Only the latter is truly compatible with Semantic Efficacy.

I then provide two arguments for Semantic Efficacy. In Section 10.3, I describe how our scientific practices appeal to mental contents as causal

difference-makers and provide reason for inductive pessimism about eliminating semantic generalizations from our scientific practice. I then explain why we should not consider this to be merely a practical difficulty: the scale at which mental contents are causally relevant is different from the scale at which intracranial states are causally relevant, such that descriptions in terms of mental content tend to cross-cut and extend beyond intracranial descriptions. By parity with other special sciences, we should accept that semantic generalizations involve difference-makers that can only be relationally (and so not intracranially) specified.

In Section 10.4, I provide an evolutionary argument that helps us to understand why we might expect mental contents to be difference-makers at the explanatory scale we find them. Continuous with explanations in evolutionary biology, such as those involving homeostasis, a relatively "higher-level" property can play a causal role not played by any particular underlying intra-organismic mechanism, but instead emerges from the coordinated execution of such mechanisms depending sensitively on features of the body and environment. These whole-organism capabilities can be selected for and refined over the course of evolution if their causal roles are adaptive. I suggest that intentional mental states and activities could have been selected for in precisely this way: instead of relying on the adequacy of homomorphisms between internal representations and action-relevant environmental features, an organism that can directly relate to what is relevant for her purposes will be more flexible, reliable, and have capacities that can generalize to other contexts and endeavors.

If my arguments are convincing, we have general reason to expand our focus in both philosophy of mind and cognitive science from what happens between the ears to understanding *systems* of embodied beings in environments. We should seek to understand how our reasoning and other mental processes serve to relate us to what is important for our needs and goals, and we should investigate how information-processing and other intracranial activity could be sensitively related to the body and environment in order to generate mental processes that constitutively serve these functions.

## 10.2 Semantic efficacy and classical cognition

One form of the Classical Cognition thesis is *computationalism*: the view that mental states and processes are computational states and processes. While there are many other varieties of Classical Cognition, it will be useful to start the discussion here. Classical Cognition holds that all mental processes can be specified *intracranially*. While adequate specifications of processes in purely neural or chemical terms would be non-semantic, it is often thought that computational accounts do respect the causal efficacy of mental content. This is false. Understanding why

will help us see why Classical Cognition as a general position (given Semantic Externalism) is incompatible with Semantic Efficacy.

I argue that computationalism and Semantic Efficacy are incompatible in Miracchi (2020).[4] Here I review points of that paper for the purposes of examining the more general Classical Cognition hypothesis.

Computational explanations in the cognitive sciences derive their explanatory force precisely from distinguishing the facts that determine content (often largely extra-cranial) from those that are causally efficacious (intracranial formal properties). This is a key part of the explanatory strategy: a computational process by definition is one that can be specified in terms of inputs, outputs, and formal transitions between them (Hillis, 1998). If mental processes are intracranial computational processes, then we can greatly simplify the problem of naturalistically characterizing them because we can prescind from environmental causal interactions (Egan, 2014). At the same time, because computational representations *have* semantic properties due to their relationships to extracranial features, computational theories can explain the predictive power of causal generalizations that appeal to mental content:

> Computers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties. … In computer design, causal role is brought into phase with content by exploiting parallelisms between the syntax of a symbol and its semantics.
>
> (Fodor, 1987, p. 19)

The explanatory strategy of taking mental processes to be identical to computational processes is inherently *internalist*: one explains the causal properties of mental processes by specifying intracranial computational mechanisms and then uses homomorphisms between these causally efficacious non-semantic properties and semantic properties in order to explain why the system behaves in accordance with relational semantic generalizations (Gallistel & King, 2009). This *internalist explanatory strategy* commits to the causal inefficacy of mental content on principle: the project of trying to specify homomorphisms between intracranial representations and semantic contents to vindicate our semantic generalizations *is* the project of trying to find internal proxies for what are (partially) externally specified mental states. (See also Haugeland, 1998, p. 237.)

Although Classical Cognition theorists differ on how much of the computationalist approach and related commitments they are willing to endorse (such as maintaining the importance of neurological properties in specifying mental kinds (e.g., Bickle, 2003) or rejecting a commitment to a Language of Thought (e.g., Horgan & Tienson, 1996)), a commitment to mental states as intracranial representations, and mental

processes as transitions between them, is characteristic of most mainstream cognitive science (see, e.g., Adams & Aizawa, 2001). And, when we put the issues in this light, it's not hard to see why Classical Cognition has so many proponents. By trading these relational characterizations of mental states for intracranial ones, we can greatly simplify the project of explaining how mental processes work and how they are part of the natural world. Who wouldn't want to have their proverbial cake and eat it too? If the internalist explanatory strategy could demonstrate sufficient promise in both vindicating semantic generalizations and naturalistically describing mental processes, it would be difficult to reject.

It is plausibly the internalist explanatory strategy that is most characteristic of Classical Cognitive Science, because it accounts for both the importance of mental representations to the paradigm and theorists' resistance to specifications of mental processes that involve extracranial elements, even where these are heavily representational (e.g., Clark, 2008; Clark & Chalmers, 1998). A convincing argument for Embodied Cognition should engage with the internalist explanatory strategy directly and explain why such a prima facie attractive view should be rejected. I think it should be rejected because (i) it actually fails to vindicate the explanatory power of our semantic generalizations, and (ii) it *oversimplifies* the problem of describing the nature and grounds of mental processes, thereby distorting the phenomena under study.

Why does Classical Cognition fail to vindicate our semantic generalizations? Causal generalizations purport not just to be predictive, but to specify the factors that causally *make a difference* to the explananda.[5] Just as we don't properly explain the advent of a storm coming by appeal to a barometer reading even though this is an excellent predictor, we don't properly explain mental events or behaviosr by citing well-correlated but non-causally efficacious features.

A computational account of mental processes supposes that one can specify their causal features independently of whatever extra-cranial features determine the contents of mental representations. If one can do that, *and* then one supposes that contents are causally efficacious, one commits to these content-determining extra-cranial features being both causally unnecessary – because irrelevant for the *ex hypothesi* sufficient intracranial description – and causally necessary, because crucial for content-determination. This is incoherent if not a downright contradiction.[6]

It should be clear that this problem extends to the internalist explanatory strategy more generally. Any account that accepts both Semantic Externalism and that the causally efficacious properties of mental processes are wholly determined by intracranial features will not be able to vindicate Semantic Efficacy. The extent to which Classical Cognition provides an error theory for our explanatory practices invoking mental contents is thus hard to overstate. If we accept the internalist explanatory strategy, then *what* we think never really matters: we are *always* citing

merely systematically correlated factors when we purport to explain by citing mental contents. We never really do anything out of love *for a person*, or because we think *that it's the right thing to do*.

I think that this is sufficient reason to cast Classical Cognitive science into serious doubt. Absent compelling empirical reason to reject Semantic Efficacy, our scientific and philosophical research programs should seek to respect and explain it. However, those already swayed by the explanatory promise of the internalist strategy may not be convinced. Accepting both Semantic Efficacy and Semantic Externalism seriously complicates the project of explaining how mental processes obtain in virtue of more fundamental processes, requiring us to integrate descriptions of intracranial processes with descriptions of bodily and environmental contributions. If the internalist explanatory strategy could make good on the rest of its promises, many would be willing to settle for Classical Cognition's simpler approach, accepting that we are simply wrong in attributing causal efficacy to mental content. Hence the rest of this paper, which offers a general defense of Semantic Efficacy intended for those interested in scientific theorizing about mental processes.

*A note before continuing:* While I will offer no suggestion here for understanding precisely how bodily and external factors play a role in metaphysically determining mental contents and content-sensitive mental processes, I think there is a lot one can say. I have elsewhere argued that we can give up the simpler computationalist strategy in a way that can be made methodologically rigorous and tractable (Miracchi, 2017a, 2019), even if the project is more difficult and the resulting theory more complex. I have also argued that perception can have accuracy conditions and rationalize beliefs even if it is not a representational state (Miracchi, 2017b), and that knowledge and belief are best understood as kinds of relational performances by agents (Miracchi, 2015, forthcoming; see also Sosa, 2007). While there is clearly more work to do, there are other games in town. (Indeed, I think the prospects for non-representational Embodied Cognition approaches are very bright, but arguing for that is a task for another day.)

## 10.3 Semantic efficacy and the special sciences

Even if appeal to the prevalence of commonsense causal explanations invoking mental contents is not convincing, we should accept Semantic Efficacy because of the prevalence, and plausible ineliminability, of this practice in the special sciences. Generally, when our scientific practices require the inclusion of certain entities and processes in our ontologies, that is grounds to think they exist as specified (Strevens, 2017). So, if our semantic causal generalizations are indeed ineliminable from scientific practice, we should accept that mental contents are causally efficacious and genuinely explanatory. I will now argue first that we have inductive

grounds to think that semantic causal generalizations are ineliminable from scientific practice, and second that semantic causal generalizations specify higher-level difference-makers that are unlikely to be supplanted by intracranial features. (Here I just take a state or process to be relatively *higher-level* to the states and processes it obtains in virtue of, taking those states and processes to be relatively *lower-level*.)

Appeal to mental contents in causal explanations is widespread in the special sciences. Apart from cognitive science (more on that below), social psychology (just look at any introductory textbook, e.g., Hewstone, Stroebe, & Jonas, 2012), economics (Dietrich & List, 2016), linguistics (especially pragmatics and sociolinguistics (Meyerhoff, 2011)), and animal behavior (Allen & Bekoff, 1997) all make essential appeal to mental contents in explaining mental processes and/or purposeful behavior, as well as other phenomena, like stock market activity.

Let us briefly examine some attempts to eliminate content-involving descriptions of behavior or cognitive processes. The traditional and clearest example of this is, of course, behaviorism. Behaviorism has largely fallen out of favor in most of psychology and cognitive science due to its inability to explain animal and human cognitive processes (see Graham (2019) for an overview). While it does retain some proponents in animal ethology (Kennedy, 1992) and economics (Gul & Pesendorfer, 2008),[7] this is increasingly becoming a minority position. As these sciences have developed, we have ever more evidence that perspicuous descriptions of animal behavior cross-cut non-agential behavioral characterizations (Keeley, 2004), and that we must include details about human psychology if we are to make appropriate predictions in economics (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974; Dietrich & List, 2016).

Much more popular and widespread is the attempt to eliminate, at least in principle, appeal to contents in our causal explanations by characterizing mental processes in terms of computational or neural intracranial features. There is often an assumption that appeal to mental contents can in principle be eliminated in favor of non-mentally characterized "behavior", neural or purely formal (computational) features of brains, or both. However, this has not been substantially borne out in practice.

Despite the popularity and interest in incorporating neuroscientific approaches in these disciplines, they have not shown promise in supplanting our semantic generalizations. Instead, the focus remains on using experimental neuroscience to gain some insight into relevant mechanisms and/or treatments for illnesses (see, e.g., Stanley & Adolphs (2013) for an optimistic overview of social neuroscience). Indeed, some prominent researchers are even arguing that the reverse is true – that neuroscience must pay more attention to behavior – much of which is characterized intentionally – if it itself is to be done properly (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017).

When we look at claims of intracranial explanations of cognition, the strongest claims we tend to find are that manipulation of some intracranial variable (cellular, hormonal, etc.) can make a difference to cognition-involving behavior. I do not contest this: what is required in order to reject Semantic Efficacy is the stronger claim that the *full* cognitive process can be intracranially described, something that is much more difficult to defend and rarely rigorously argued for. Instead, there is often a slide in relevant discussions between the weaker (still very interesting and important) claim that intracranial mechanisms can be shown to make a difference to cognition and the much more substantive thesis that cognition itself can be intracranially described. For example, John Bickle takes himself to have established the claim that:

> Current reductionistic neuroscience now provides causal-mechanistic explanations of behaviors routinely taken to indicate cognitive functions…
>
> (Bickle, 2015, p. 307)

In support of this claim, Bickle provides the example of research designed to show that the alpha isoform of calmodulin kinase II (α-CaMKII) is causally relevant to spatial memory. Even if his argument goes through, what he actually establishes is only that α-CaMKII is a difference-maker for spatial memory. The experiments adduced do not provide a theory of the whole mechanism that would render appeal to memory contentfully described as unnecessary in the explanation of mice behavior. Although Bickle himself admits this –

> Establishing that a molecule like α-CaMKII is part of the causal mechanism for both synaptic LTP and spatial memory does not tell us how it is so causally linked – about what other mechanisms mediate α-CaMKII activity in hippocampal neurons and trained behavior in the Morris water maze, for example. That requires more connection experiments investigating related causal hypotheses…
>
> (Bickle, 2015, p. 308)

– he does not address the gap between what he actually establishes and the claim that he has provided an intracranial causal-mechanistic account of cognitive processes. Given the admitted difficulty of even establishing that α-CaMKII is relevant to spatial memory, we cannot assume that our contentful characterizations of cognitive processes can be successfully made redundant anytime soon.

Perhaps the most widespread defense of the commitment that contentful characterizations of mental processes can be eliminated from our scientific practice is the research program of computationally or information-theoretically describing processes. On such a view, appeal

to mental contents in cognitive science is merely a shorthand for, or a sketch or "gloss" on, causal processes that can be fully specified neurologically or computationally (see Egan, 2014).

It is often difficult to evaluate the extent to which appeal to mental contents can be excised from theories in cognitive science without loss of accuracy and explanatory power, and this question is rarely explicitly addressed empirically, with researchers often leaving the elimination of contentful characterizations of mental processes as promissory or moving back and forth between contentful and formal characterizations of processes. Consider for example Baddeley's (2010) discussion of item storage in working memory. Sometimes items are more plausibly described formally (such as in phonologically distinct word items), but sometimes they are described semantically, in ways that do not indicate a clear computational specification, let alone an algorithmic implementation:

> The capacity to remember and repeat a string of unrelated words is about five items, but if they comprise a meaningful sentence, the span is around 15 words, reflecting a contribution from grammar and meaning, both depending on different aspects of long-term memory.
>
> (Baddeley, 2010, p. R140)

It is by no means clear that the contribution of meaning to length of word recall can be eliminated in favor of a formal characterization. Thus, despite the commitments of such an approach, the extent to which these research programs make it plausible that contentful descriptions of mental processes can even in principle be eliminated is deeply unclear.

We have further reason to doubt that such approaches will make good on their promises by examining research done in artificial intelligence and robotics, which largely operates on the commitment that mental processes can be computationally, and so non-semantically, characterized. Traditional classical "Good Old-Fashioned AI" operating exclusively within this model is widely agreed to have failed, suffering both from problems of computational intractability and from the so-called "Frame Problem" for AI (see Russell & Norvig, 2014).[8]

The fragility of these systems, i.e., the inability of researchers to imbue these computational systems with the right kinds of programming so that they behave robustly in accordance with the kinds of regularities that semantic generalizations specify, suggests that eliminating contentful characterizations of mental processes in favor of formal characterizations will be much more difficult than is often supposed.[9]

Newer research programs to some extent liberalize classical assumptions, but typically do not significantly depart from them (see Miracchi (2019) for discussion). And despite the impressive advances of the last few decades of AI and robotics research in building specialized systems,

we have experienced severe limitations in generalizing these successes to more sophisticated tasks in real-world environments. We are very far from developing artificial agents who behave similarly to how minded animals and humans do – i.e., in the kinds of ways that we semantically describe: ways that would indicate acting on one's desires, understanding one's surroundings, or even simply carrying out one's plans (e.g., see Sofge (2015) for an overview of the disappointing failures at the 2015 DARPA Robotics Challenge).

The argument I wish to provide here is not merely inductive (although it is partially that). When we look at our practice of providing semantic descriptions of mental processes in the special sciences, we see robust *difference-makers* described in semantic terms, which can be intervened on and manipulated *as such* to make an impact. Consider the toy example Dietrich and List (2016) use in arguing for the ineliminability of mental generalizations in economics:

> Consider, for example, how you would explain a cat's appearance in the kitchen when the owner is preparing some food. You could either try (and in reality fail) to understand the cat's neurophysiological processes which begin with (i) some sensory stimuli, then (ii) trigger some complicated neural responses, and finally (iii) activate the cat's muscles so as to put it on a trajectory towards the kitchen. Or you could ascribe to the cat (i) the belief that there is food available in the kitchen, and (ii) the desire to eat, so that (iii) it is rational for the cat to go to the kitchen. It should be evident that the second explanation is both simpler and more illuminating, offering much greater predictive power. The belief-desire explanation can easily be adjusted, for example, if conditions change. If you give the cat some visible or smellable evidence that food will be available in the living room rather than the kitchen, you can predict that it will update its beliefs and go to the living room instead. By contrast, one cannot even begin to imagine the informational overload that would be involved in adjusting the neurophysiological explanation to accommodate this change.
>
> (Dietrich & List, 2016, p. 275)[10]

What Dietrich and List's example helps to make clear (although they do not explicitly draw this point out) is that the explanatory power of appealing to mental contents doesn't derive purely from its simplicity compared to any neuro-behavioral account that might replace it, but from the understanding, it gives us of how to intervene at the level of mental content to manipulate outcomes, both in theory and practice. A better understanding of the relation between manipulability and causation is one of the biggest advances in the last 20 years of philosophical work on causation, and it is now widely accepted that causal explanations

include a kind of asymmetry that encodes (or at least entails) knowledge about how the world would change under different interventions (e.g., Strevens, 2008; Woodward, 2003; see Woodward and Ross (2021) for an overview). The prevalence of semantic generalizations in the special sciences, together with their resistance to elimination, suggests that semantically characterized mental states are important difference-makers in their own right.

Now, the Classical approach will hold that in every case what one intervenes on are really intracranial features of mental representations; but because these features systematically co-vary with mental contents, we can manipulate the world by acting *as if* we were manipulating mental events contentfully described. Despite the inductive pessimism I motivated above about this research program, there are further reasons to resist this claim.

Note first that, throughout the special sciences, higher-level kinds are often difference-makers where their lower-level constituents are not. See, e.g., Strevens (2004, 2008) for discussion of the irrelevance of individual molecular activity and other lower-level properties of containers to the truth of Boyle's law. The difference-making relationship between pressure and volume is not one that can be described in terms of lower-level features. A defense of the Classical approach would give us reason to think that semantic generalizations are *special* in the sense that they are high-level generalizations that are highly useful for causally influencing (animals and people in) our environments but nevertheless the real difference-makers are sub-components of the systems.

Let us consider what it would take for this to be true. Content-involving descriptions of mental processes are highly relational, often cross-cutting any plausible candidates for intracranial non-semantic descriptions (Dennett, 1987a; Peacocke, 1994, e.g., p. 45; Miracchi, 2020). Consider, for example, *A* inviting *B* to dinner. This typically results in *B* inferring that *A* has invited her to dinner. Invitations can vary widely in their physical forms, in ways that depend on context, prior conversation, etc. (Not all invitations take the explicit form "Would you like to have dinner?") An account of the processes by which one makes such an inference that eliminated appeal to mental contents would have to say something about the various sensorial stimulation and subsequent information-processing in context that groups them together, resulting in a representation, or a member of a class of representations, with the content that one has been invited to dinner. Ditto, mutatis mutandis, for the effects of such knowledge, as well as any of its other potential causes.

But *ex hypothesi* whatever basis one would have for the grouping would not be semantic. It could at most merely represent what these external events have in common *qua* invitations. If we were to change the contexts or inessential features of the invitational acts, we could then bring the internal representations *out of* systematic correlation with the

external features they represent. In this way, representational systems are highly brittle. Whereas mental generalizations are relational and purport to relate the agent directly to features of her environment (e.g., knowing that you've been invited to dinner because you've been told) the internalist explanatory strategy provides generalizations that only hold in the conditions for which there is a homomorphism between internal representations and semantic structure. The Classical theorist is then on the hook for *ensuring* that internal representations are rich enough to secure the requisite homomorphism with these relational variables across the range of environmental conditions the subject might encounter. This is impossible to do generally, given the contingent relationship between intracranial structure and external environmental features.

In practice, it is likely to be difficult even for restricted systems and environments. In order to preserve homomorphisms across changes in context, the theorist would need to take seriously into account the various contingent relationships between features of internal representations and semantic generalizations. Getting the right computational processes to be triggered in response to stimuli and motor output in order to ensure the effectiveness of the interventions specified by semantic generalizations would require such attention to the relational facts that determine these homomorphisms that the resulting theory would no longer vindicate the internalist explanatory strategy. It would not account for mental processes largely in abstraction from extra-cranial features.

Let us regroup. The Classical Cognition theorist purports to simplify the project of explaining mental processes by appeal to intracranial mental processes whose non-semantic properties are appropriately homomorphic to the relational semantic characterizations. But when we acknowledge the robust explanatory power of semantic generalizations, enabling extensive prediction and difference-making, it becomes clear that any execution of the internalist explanatory strategy that could do justice to such a network of generalizations would have to specify features of the internal system in such a detailed and precise way so as to explain how homomorphisms are preserved even across wide changes in (intracranial and extracranial) context. Such an approach no longer looks very simple and plausibly includes exactly the kind of careful attention to bodily and environmental features that Embodied Cognition proponents advocate and Classical Cognition proponents hope to avoid. We thus lose the central motivation for appealing to internal representations in our theorizing about mental processes. Appeal to internal representations no longer simplifies our study of mental processes or the prospect of naturalizing them, because regardless of whether we appeal to them we must pay careful attention to the way intracranial information processing is sensitive to features of the body and environment.

The more plausible approach, in my view, is to take the relational, semantic characterizations of mental processes at face value: that is,

to accept Semantic Efficacy. Only an account that respects Semantic Efficacy can respect the explanatory roles of semantic generalizations in both everyday reasoning and scientific practice. Semantic generalizations are true in the first instance of embodied agents that take attitudes (perception, belief, desire, intention, etc.) towards features of their environments. Mental processes, therefore, should not be understood in terms of transitions between internal representations, but instead in terms of transitions between relational states or activities of embodied agents. A study of *how* mental processes work would seek to describe these transitions in contentful, relational terms. A study of *in virtue of what* we make these transitions would seek to understand how intracranial processes, in relation to body and environment, gives rise to these robust relational regularities.

Note that nothing I have said here rules out the (in my mind very plausible) view that neural information processing of the sort Classical Cognition theorists study is a large part of the explanatory story of how content-involving mental states and processes are generated. It is just that these representations will not be mental representations, and the processes involving them will not be mental processes. They will be sub-personal.[11]

Lack of clarity about this issue has generated confusion and difficulty on both sides. While Classical Cognition theorists make the mistake of thinking that genuine causal explanations of mental events must involve non-mentally specifiable intracranial states and processes, Embodied Cognition theorists often make the mistake of thinking that the denial of the classical position requires them to reject *any* appeal to neural or computational representations in specifying the lower-level kinds and processes that give rise to mental processes (e.g., Beer, 2003; Brooks, 1991; Campbell, 2002; Chemero, 2009; Gibson, 1979; Varela, Rosch, & Thompson, 1991, ch. 7; van Gelder, 1995, 1998; and see Shapiro, 2011, esp. ch. 5, for discussion). What should be denied is that the mental processes *are* intracranial representational processes, not that there are no intracranial representational processes relevant to cognition.[12]

## 10.4 An evolutionary argument for semantic efficacy

Now I will provide an evolutionary motivation for the previous conclusion: namely that semantic descriptions of mental processes are the only accurate descriptions we are going to be able to find for the vast majority of cases. Recall that semantic generalizations are highly relational: they relate the agent to spatiotemporally distant features of her environment. Whereas Classical Cognition holds that such generalizations are always replaceable with intracranial generalizations, we have reason to think that causal processes and regularities with this wider scope can be, and

in fact, were, selected for. Evolution can do better than representation: it can make us genuinely sensitive to what matters for us.

Let's consider an analogous issue in biology. As Holly Andersen (2013) points out, biological systems regularly violate what is called "Causal Faithfulness": they often have precisely counterbalancing causal mechanisms, so that two factors that are actually causally related are statistically independent, at least under a wide range of ecological and experimental conditions. There's good reason why nature does this. As Andersen points out:

> The tendency for evolved systems like populations, individual organisms, ecosystems, and the brain to involve precisely balanced causal relationships can be easily explained by the role these balanced relationships play in maintaining various equilibrium states (see, e.g., Mitchell (2003, 2008)). Furthermore, the mechanisms by which organisms maintain internal equilibrium with respect to a huge variety of states need to be flexible. They need to not simply maintain a static equilibrium but maintain it against dynamic perturbation from the outside. This means that many mechanisms for equilibrium maintenance can maintain a fixed internal state over some range of values in other variables. Thus, a system that survives because of its capacity to maintain stability in the face of changing causal parameters or variable values will be likely to display CF-violating causal relationships and will also violate the stronger condition of causal stability.
>
> (Andersen, 2013, p. 678)

(See also Kitano (2004) for discussions of counterbalancing mechanisms in biological systems.) Here Andersen suggests that systems violating Faithfulness are often selected for by evolution because precisely coordinated constitutive lower-level mechanisms can preserve fitness-enhancing higher-level states, like homeostasis.

This raises two distinct issues which are worth explicitly articulating. First, we understand how the higher-level state or process obtains in virtue of lower-level states and processes by adequately specifying how these various internal mechanisms are counterbalanced in ways that are highly sensitive to environmental factors, so as to robustly preserve the higher-level state across changing environmental conditions. Thus the higher-level state or behavior can only be understood as constituted by the lower-level complex of internal mechanism-environment interactions. There is no underlying internal mechanism, or system of mechanisms, that corresponds to the higher-level state or behavior.[13]

Second, causal generalizations that *appeal* to equilibrium states as explanans cannot be eliminated in favor of generalizations involving underlying mechanisms. Consider thermoregulation. It is because an

organism has the ability to stay at roughly the same temperature across a range of environmental conditions that its enzymes can function well, cell permeability remains at optimal levels, etc. (Freeman, 2005). No underlying mechanism has those properties. Moreover, in many cases, we cannot intervene on the property of thermoregulating by intervening on underlying mechanisms because counterbalancing mechanisms would be triggered, preserving higher-level functioning. (This is the violation of "Faithfulness" Andersen describes above.) The mechanisms that give rise to homeostasis are not explanatory of the same phenomena that homeostasis is.

This point generalizes well beyond equilibrium states, which are Andersen's focus. The coordination of lower-level mechanisms sensitively with the environment will not just be useful for maintaining equilibrium states, but will be useful for giving rise to robust whole-organism adaptive states and behaviors generally: finding food, avoiding predators, mating, caring for children and sick relatives, etc. There will be a selective pressure to move from the ability to respond to more proximal states that *indicate* the presence of things relevant for action in the agent's environment, towards organisms that can directly and robustly respond to those relevant things *themselves*, i.e., who treat other animals as predators, prey, mates, young in how they interact with them. In continuity with biological processes, evolution may have selected for organisms that had increasingly relational agential abilities *per se*. We can understand, then, how animals' purposefully interacting with other organisms and features of their environments could have developed as the most fundamental form of original intentionality.[14]

If this were true, then by parity with other biological cases like homeostasis, we would expect lower-level intracranial mechanisms to be coordinated with one another in ways that are quite complicated, and subtly sensitive to changes in body and environment across a wide range of circumstances to produce increasingly robust and extensive relational organism-environment activity.

We can generalize this suggestion to include even the highest forms of cognition which do not have specific adaptive benefits, and which are thought to be most plausibly "internal", or symbolic, such as curiosity, counterfactual reasoning, and the ability to plan and reason in creative ways. Louise Barrett and others refer to these capacities as "long leash" fitness enhancers (Barrett, 2011) because they do not produce a specific adaptive behavior or trait, but rather tend to be selected for because organisms with these capacities can more effectively cope with novelty, change, and a broader range of conditions. Such animals can more effectively find food, avoid predators, protect themselves from inclement weather, develop fitness-enhancing social relationships, etc. While these cognitive abilities are selected for, plausibly they are not selected for any fitness-enhancing behavior in particular. Instead, what they do is enable

humans and non-human animals to relate to their environments in ever more abstract, sophisticated, and spatiotemporally extended ways, promoting survival and reproduction more generally.

Whatever relational facts are important for determining mental content and these more abstract content-involving mental processes, it is likely that we must take them into account in explaining their natural bases. First, it is likely that these higher-level cognitive processes are significantly metaphysically dependent on lower-level cognitive processes, which themselves are dependent on the body and environment. Second, these more abstract processes are, even Classical Cognition theorists would admit (aside from deductive reasoning perhaps), more resistant to formal treatment: curiosity, creativity, and play are mental activities where mental content seems deeply relevant. This gives us reason to think that body and environment will be even more relevant, not less, regardless of whether overt behavior is involved.

Some support for this view comes from Paul Cisek's (2019) "phylogenetic refinement" approach, where he adduces both phylogenetic and neuroscientific evidence in order to reject the traditional perception-cognition-action linear cognitive architecture favored by Classical Cognitive Science in favor of an action-organized architecture that reflects animal evolution from less to more sophisticated animal behaviors, involving sensorimotor loops relating both internal and external factors.[15] According to Cisek, choosing among behaviors is not attributed to the internal processing of mental representations, but results from the evolutionary development of an organism that tends to make the appropriate trade-offs given its context and condition. More sophisticated and abstract forms of cognition, he suggests, are likely to be understood continuously with these more basic behaviors, plans, and decisions:

> It may be that a serial cognitive "virtual machine" appeared in human brains atop their inherited architecture for flexible primate behavior (Block, 1995). However, I believe it unlikely that such a major redesigning of the functional architecture could have happened within the last few million years, after what had been nearly a billion years of relatively continuous differentiation and specialization of closed-loop feedback control systems. It seems more promising to consider how that architecture of nested feedback control, which has been extending further and further into the world for so long, might have just kept extending into increasingly abstract domains of interaction (Hendriks-Jansen, 1996; Pezzulo & Cisek, 2016).
>
> (Cisek, 2019, p. 15)

This kind of action-driven approach is plausibly more parsimonious and empirically grounded than the Classical approach: it is likely that in understanding the evolution of the brain we must pay significant

attention to how parts of the brain are selected-for because of the roles they play in adaptive behavior, and that these are going to involve the kind of metaphysical dependence on body and environment we have been considering. Embracing an Embodied Cognition approach would not *rule out* the possibility of discovering that certain more abstract mental processes could be characterized wholly internally; it would just not assume it as a methodological commitment of scientific practice. By treating these more abstract cognitive functions as developments of a deeply embodied and embedded neural architecture that generates robust relational properties at the agent level, we can investigate empirically what kinds of internal architecture and representations are necessary for the production of more sophisticated and abstract cognition.

## 10.5 Conclusion

I hope to have made a compelling case for endorsing Semantic Efficacy, and therefore Embodied Cognition. Upon closer scrutiny, Classical Cognitive science does not make good on its promises either to vindicate our practice of providing causal explanations by appeal to semantically characterized mental states nor does it make a convincing case for being able to simplify the projects of describing mental processes and in virtue of what they obtain by prescinding from bodily and environmental involvement. The argument does not depend on details about the nature of mental states and processes, the kind of metaphysical dependence they have on intracranial features, body, or environment, or the way in which mental content is determined by external features. This keeps the debate between Classical and Embodied Cognition theorists at a suitably high level of generality, helping us to see the broader structural issues between the two approaches without getting bogged down in specific theories.

The defense of Semantic Efficacy I have provided here has not involved special pleading on behalf of the mental. Rather, I have attempted to place our semantic generalizations as part of a general practice in the special sciences of explaining by specifying higher-level difference-makers. I have argued for Semantic Efficacy by urging parity of treatment in how these higher-level difference-makers are thought to relate to lower-level processes and theorizing about how they might have evolved. As a result, I hope to have defused much concern that one might have over the naturalistic acceptability of Semantic Efficacy, without offering a metaphysics of mental states or contents.

If my arguments have been convincing, cognitive science and philosophy of mind should substantially broaden their theoretical and explanatory purview. No longer can we assume that mental processes can be placed between the ears. Instead, we should embrace the messier, but hopefully much more realistic, problem of understanding how our

cognitive capacities serve to relate us to the features of our world that are important for satisfying our needs and accomplishing our goals. In seeking to understand the natural bases of mental processes we must start with more fundamental forms of cognition and investigate how information-processing and other intracranial mechanisms could be orchestrated so as to produce robust relational agential activity by whole organisms across a range of conditions. Only with an understanding of these more basic cases can we hope to understand the natural bases of more sophisticated forms of cognition.

Lastly, lest the claims made about mental processes here be misunderstood: the claim that mental processes are inherently relational does not commit us to the idea that *the thinker* extends beyond her body into her environment. Just like other agent-environment processes, for example, playing catch, the agent does not extend as far as what she does. The agent is wholly in her own body, on my view, but cognition gives her grasp of much beyond herself.

## Notes

1. See Shapiro (2011), esp. ch. 6 for a useful distinction between this embodied cognition thesis (which he calls "The Constitution Hypothesis") and other related theses.
2. While I will sometimes use the term "constitution" or "grounds" in the discussion, this is not intended to change the fundamental form of the argument. Please feel free to substitute your favorite metaphysical relation.
3. Although there are sophisticated attempts to do without this (e.g., Loar, 1987), it proves very difficult to do and is itself somewhat mysterious. See Miracchi (2017c) for some further discussion.
4. I have changed my name. Please cite this and future work under "Titus, Lisa Miracchi" or "Titus, Lisa, M." When discussing my previous work inline, please refer to me as "Titus, (née Miracchi)."
5. See also Woodward (2003) and Strevens (2004).
6. Rescorla (2014) argues that on an interventionist approach one should accept both the sufficiency of formal properties for explaining mental processes and Semantic Efficacy as a kind of non-problematic overdetermination. In future work I hope to show why this argument rests on problematic *metaphysical backtracking*, where the purportedly justificatory counterfactuals only hold because the features that are causally relevant for the mental process are metaphysically relevant to mental content. For our purposes here it is enough to note that we must take care when inferring causal or metaphysical relationships from counterfactuals. The fact that Classical Cognition relies on the inessentiality of content to causal explanations should make us skeptical that counterfactuals involving semantic contents reveal causal relationships.
7. Proponents in psychology and cognitive science include Ivan Pavlov (1906) and B.F. Skinner (1953).
8. I distinguish the more general Frame Problem from the more specific logical problems on which we have seen considerable progress (Shanahan, 2016).
9. I argue for this claim in detail in Miracchi (2020).

10. See also Dietrich & List's discussion of functional generalizations (2016, p. 275).
11. In Miracchi (2020) I advocate a generative approach to the relationship between computation and mental and behavioral processes that are semantically characterized, contrasting it with the view that mental processes are *identical to* computational processes. Elsewhere (Miracchi, 2017a), I motivate a general separation between causal projects (understanding what makes a causal difference to what) and generative projects (understanding what gives rise to, or makes a generative difference to, what). Classical cognitive science, in seeking to identify mental processes with computational processes, collapses these projects together, thereby inhibiting both.
12. See also Miracchi (2017b) for discussion.
13. See Hurley (1998) for a precursor of this kind of Embodied Cognition approach.
14. See Miracchi (2017b, 2019) for more discussion of this claim. See Burge (2010) and Godfrey-Smith (2016) for related but weaker positions which still prioritize the development of representational capacities as crucial for intentionality.
15. Although his account is still largely within the representationalist paradigm, he is careful to point out that on his view representations will be more "pragmatic", i.e., inherently tied to the production of action, and that there will be less of a focus on intracranial processes of decision-making and planning to produce action, but instead on competition between different functional circuits to produce adaptive behavior. I invite the reader to see the current contribution as a way to simplify and extend this kind of approach, as well as a proposal for how to include mental processes like planning and decision-making in such a research paradigm by releasing ourselves from Classical conceptions of such processes.

## References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, *14*(1), 43–64.

Aizawa, K. (2007). Understanding the embodiment of perception. *Journal of Philosophy*, *104*(1), 5–25.

Allen, C., & Bekoff, M. (1997). *Species of mind: The philosophy and biology of cognitive ethology*. Cambridge, MA: MIT Press.

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, *80*, 672–673.

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140.

Barrett, L. (2011). *Beyond the brain: How body and environment shape animal and human minds*. Princeton, NJ: Princeton University Press.

Beer, R. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11*, 209–243.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bickle, J. (2015). Marr and reductionism. *Topics in Cognitive Science*, *7*, 299–311.

Block, N. (1995). The mind as the software of the brain. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (pp. 377–425). Cambridge: MIT Press.

Block, N. (2005). Action in perception by Alva Noë. *Journal of Philosophy*, *102*(5), 259–272.
Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.
Burge, T. (2010). *Origins of objectivity*. New York: Oxford University Press.
Campbell, J. (2002). *Reference and consciousness*. New York: Oxford University Press.
Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, and Psychophysics*, *81*(7), 2265–2287.
Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.
Dennett, D. (1987a). Cognitive wheels: The frame problem of AI. In Z. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence* (pp. 41–64). New York: Ablex.
Dennett, D. (1987b). *The intentional stance*. Cambridge, MA: MIT Press.
Dietrich, F., & List, C. (2016). Mentalism versus behaviorism in economics: A philosophy-of-science perspective. *Economics and Philosophy*, *32*, 249–281.
Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115–135.
Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
Freeman, S. (2005). *Biological science* (2nd ed.). Hoboken, NJ: Pearson Prentice Hall.
Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. New York: Wiley/Blackwell.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.
Godfrey-Smith, P. (2016). *Other minds*. New York: Farrar, Straus, and Giroux.
Graham, G. (2019). Behaviorism. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Retrieved from https://plato.stanford.edu/archives/spr2019/entries/behaviorism/
Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. In A. Caplin, & A. Schotter (Eds.), *The foundations of positive and normative economics* (pp. 3–39). New York: Oxford University Press.
Hendriks-Jansen, H. (1996). *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought* (p. 8). Cambridge: MIT Press.
Hewstone, M., Stroebe, W., & Jonas, K. (Eds.). (2012). *An introduction to social psychology* (5th ed.). Chichester, UK: BPS Blackwell and John Wiley & Sons.
Hillis, D. W. (1998). *The pattern on the stone*. New York: Basic Books.
Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology*. Cambridge, MA: MIT Press.
Hurley, S. L. (1998). *Consciousness in action*. Cambridge, MA: Harvard University Press.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.
Keeley, B. L. (2004). Anthropomorphism, primatomorphism, mammalomorphism: Understanding cross-species comparisons. *Biology and Philosophy*, *19*, 521–540.

Kennedy, J. S. (1992). *The new anthropomorphism*. Cambridge: Cambridge University Press.
Kitano, K. (2004). Biological robustness. *Nature Reviews Genetics*, *5*, 826–837.
Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490.
Loar, B. (1987). Subjective intentionality. *Philosophical Topics*, *15*(1), 89–124.
Meyerhoff, M. (2011). *Introducing sociolinguistics* (2nd ed.). New York: Routledge.
Miracchi, L. (2015). Competence to know. *Philosophical Studies*, *172*(1), 29–56.
Miracchi, L. (2017a). Generative explanation in cognitive science and the hard problem of consciousness. *Philosophical Perspectives*, *31*(1), 267–291.
Miracchi, L. (2017b). Perception first. *Journal of Philosophy*, *114*(12), 629–677.
Miracchi, L. (2017c). Perspectival externalism is the antidote to radical skepticism. *Episteme*, *14*(3), 363–379.
Miracchi, L. (2019). A competence framework for artificial intelligence research. *Philosophical Psychology*, *32*(5), 589–634.
Miracchi, L. (2020). Updating the frame problem for AI research. *Journal of Artificial Intelligence and Consciousness*,, *7*(2), 217–230.
Miracchi, L. (forthcoming). Competent perspectives and the new evil demon problem. In F. Dorsch, & J. Dutant (Eds.), *The new evil demon: New essays on knowledge, justification and rationality*. Oxford: Oxford University Press.
Mitchell, S. (2003). *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
Mitchell, S. (2008). Exporting causal knowledge in evolutionary and developmental biology. *Philosophy of Science*, *75*(5), 697–706.
Noë, A. (2005). *Action in perception*. Cambridge, MA: MIT Press.
Pavlov, I. (1906). The scientific investigation of the psychical faculties or processes in the higher animals. *Science*, *24*(620), 613–619.
Peacocke, C. (1994). Content, computation, and externalism. *Mind and Language*, *9*(3), 303–335.
Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences*, 20, 414–424.
Rescorla, M. (2014). The causal relevance of content to computation. *Philosophy and Phenomenological Research*, *88*(1), 173–208.
Russell, S., & Norvig, P. (2014). *Artificial intelligence: A modern approach* (3rd ed.). Hoboken, NJ: Pearson Prentice-Hall.
Schaffer, J. (2012). Grounding, transitivity, and contrastivity. In F. Correia, & B. Schnieder (Eds.), *Metaphysical grounding: Understanding the structure of reality* (pp. 122–138). Cambridge: Cambridge University Press.
Shanahan, M. (2016). The frame problem. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). Retrieved from https://plato.stanford.edu/archives/spr2016/entries/frame-problem/
Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.
Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
Sofge, E. (2015, July 06). The DARPA robotics challenge was a bust. *Popular Science*. https://www.popsci.com/darpa-robotics-challenge-was-bust-why-darpa-needs-try-again/

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (vol. 1). Oxford: Oxford University Press.
Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron Perspective*, *80*, 816–826.
Strevens, M. (2004). The causal and unification approaches to explanation unified – Causally. *Noûs*, *38*(1), 154–176.
Strevens, M. (2008). *Depth*. Cambridge, MA: Harvard University Press.
Strevens, M. (2017). Ontology, complexity, and compositionality. In M. H. Slater, & Z. Yudell (Eds.), *Metaphysics and the philosophy of science: New essays* (pp. 41–54). Oxford: Oxford University Press.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.
van Gelder, T. (1995). What might cognition be if not computation. *Journal of Philosophy*, *92*(7), 345–381.
van Gelder, T. (1998). The dynamical hypothesis. *Behavioral and Brain Sciences*, *21*(5), 615–28.
Varela, F. J., Rosch, E., & Thompson, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.
Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.
Woodward, J., & Ross, L. (2021). Scientific explanation. In E.N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). Retrieved from https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

# 11 Two-Way Powers as Derivative Powers

*Andrei A. Buckareff*

> "Ontology is, as C.B. Martin liked to put it, a package deal. Ontology must be approached from the bottom up. If you have no patience for such things, you would be well advised to look elsewhere for philosophical issues to tackle"
>
> –John Heil (2017, p. 48).

## 11.1 Introduction

In *Metaphysics* Θ,[1] Aristotle introduces a distinction between rational powers and non-rational powers (1046a36–1046b4). Additionally, he distinguishes one-way powers (which are manifested in only one way in response to some activation conditions) from two-way powers (which can be manifested in either of two opposite ways) (1046b4–7). Finally, he identifies one-way powers with non-rational powers and two-way powers with rational powers.

Recently, some philosophers working on the metaphysics of agency (most of whom explicitly tip their hat to Aristotle) have emphasized the need for agents to possess ontologically irreducible two-way powers as a necessary condition not only for free will but for intentional agency, more generally. Specifically, these authors argue that agency is characterized by the manifestation of a strongly emergent *two-way power* either to do *A* or not-*A* at the time intentional agency is exercised.[2]

In this paper, I assume that the *concept* of two-way powers may be indispensable, particularly for how we represent ourselves and others as practical decision-makers. And I assume that we cannot reduce the concept of a two-way power to more basic concepts.[3] But while I assume that we cannot and should not attempt to reduce the concept of two-way powers, we can and should dispense with *ontologically irreducible* two-way powers. Specifically, I argue that such powers can be ontologically reduced to what George Molnar (2003) referred to as "derivative powers" in his work on the ontology of causal powers.[4]

I proceed as follows in this essay. First, in the interest of clarifying my target, I begin by outlining some generic features of two-way

DOI: 10.4324/9780429022579-12

powers – including some minor points of disagreement among defenders. Next, in order to motivate exploring an ontologically reductive alternative to irreducible two-way powers, I consider two worries: one is explanatory and the other is metaphysical. The explanatory worry is over the role of reasons in explaining the manifestations of two-way powers. The metaphysical concern is over substance dualism being an apparent ontological commitment of accepting two-way powers. The remainder of the essay is devoted to the presentation of an ontologically reductive account of two-way powers as derivative powers,[5] along with the consideration of two objections to my alternative.

## 11.2 Two-way powers

Suppose that the concept of intentional *agency* is best understood in terms of *settling* (Clancy, 2013; Steward, 2012). At least in the case of human agents, it will be true that an agent's capacity for intentional agency is manifested (and, hence, the agent exercises intentional agency) when an agent settles the truth of *p* by acting or refraining from acting. Therefore, the capacity to exercise intentional agency is, at least in part, the capacity to settle the truth of *p*.

I find that the most salient examples of representing ourselves as settling that suggest the need for us to conceive of ourselves as in possession of a two-way power involve making practical decisions between options. This is so since in making practical decisions we represent ourselves as possessing the power to select either of at least two alternatives. Thus, I assume in the following that it is not obvious that the concept of two-way powers is essential to understanding *all* intentional agency. But I presuppose that a tacit assumption we make about ourselves in representing ourselves as decision-makers with the power to select between options is that we are exercising a two-way power when we make practical decisions.

Regarding the nature of practical decisions, I identify them with mental actions of forming an intention. More specifically, practical decisions are actions of intention-formation directed at settling some practical uncertainty over whether to *A* or not to *A*. This understanding of practical decisions is similar to Alfred R. Mele's (2003; 2017). But while he takes practical decisions to be *momentary* mental actions whereby an agent settles some practical uncertainty, I do not add the restriction that decisions are momentary. The action can be momentary or extended. Therefore, I assume that decision-making may be identical with an action of extended deliberation or it may be a momentary mental action that is identical with what some have in mind in talking about *choosing*. The process of decision-making may or may not involve the performance of the act of choosing. What is important is that an

agent making a practical decision is performing an action of forming an intention about what to do among some options. So, in brief, practical decision-making is a mental action that is aimed at settling some practical uncertainty that involves an agent taking an active role in the acquisition of an intention (hence the agent's 'forming' an intention rather than automatically acquiring it in response to practical reasons). Importantly, again, I assume that when making decisions agents represent themselves as having a two-way power to decide between options.[6]

While I am happy to concede that agents making practical decisions represent themselves as exercising a two-way power, I am skeptical about the more general assertion that any representation of an agent settling the truth of *p* in cases other than decision-making would include an ontological commitment to irreducible two-way powers. Most contemporary proponents of two-way powers in theorizing about intentional agency have insisted that the power to settle whether *p* requires that agents possess irreducible two-way powers. Hence, they have asserted that *any* exercise of intentional agency involves the manifestation of a two-way power. For instance, Erasmus Mayr takes what he identifies as the "active powers" that are characteristic of human agents that are active in exercising agency to be "two-way powers which the agent can exercise or refrain from exercising" (2011, p. 231). And, while she denies that such two-way powers are unique to humans, Helen Steward underscores that, on her account, the exercise of the power "to settle *which* of [many] courses of action becomes the actual one" in exercising intentional agency involves the exercise of "the two-way power of agency" (Steward, 2012, p. 173). On the sort of view these authors represent, there is no agency or agent who is producing outcomes if there are no irreducible two-way powers being manifested since the manifestation of a two-way power is necessary for some behavior to involve an exercise of agency.

But what *are* two-way powers? It may first be best to say something general about all powers. I will assume that all powers are properties of objects, properties are ways objects are; and objects are either reducible to bundles of properties or irreducible substances. In particular, I assume that powers are identical with the dispositional properties of objects.[7] I assume that properties are either particulars (tropes or modes of substances) or immanent universals that are wholly present in their property-instances. So, to be more precise, the individual powers possessed by objects are either properties (understood as particular modes of objects) or particular property-instances (where properties are understood as immanent universals). Such an assumption about powers is not terribly controversial these days in the literature on the ontology of powers. But among prominent defenders of two-way powers, there is a plurality of views. While some (e.g., Kenny, 1989; Lowe, 2008; 2013a; 2013b; Mayr, 2011) identify powers with properties, others are silent

on the matter (e.g., Alvarez, 2013) or express sentiments inimical to any such identification (e.g., Steward, 2012, pp. 222–223). In defense of identifying powers with properties, I find it puzzling how a power can fail to be a property (or an instance of a property) since it is a *way* that an object is. That is, a power characterizes an object. Moreover, if two-way powers are basic powers of agents, as the proponents of two-way powers in the philosophy of agency insist, then it is hard to see how they are basic *features* of agents without being properties. Ergo, it seems that the denial of identifying powers with properties (or their instances) rests on either some confusion about what a property is or else a commitment to a view of properties on which they (or their instances) are not in space-time. I will, therefore, treat the assumption that powers are properties as relatively innocuous, being mindful of the fact that some might resist any such identification.

If we wish to make headway on understanding *two-way powers* it helps to contrast them with *one-way powers*. One-way powers are *causal* powers manifested in *one specified way* when partnered with a specific manifestation partner or constellation of partners (e.g., when there is just a neutral chlorine atom with which it interacts, a neutral sodium atom will manifest its power to lose an electron to the chlorine atom which has the power to gain the electron from the sodium atom).

While one-way powers will manifest in a specific way and only in that way when partnered with some reciprocal manifestation partners, the term 'one-way power' is deceptive. Many one-way powers may be manifested in different ways, contribute to various outcomes in a causal process, and they cooperate with other powers to bring about outcomes. So many one-way powers are described as *multi-track, pleiotropic*, and contributors to *polygenic* effects.

First, many one-way powers are *multi-track* because they are directed at a range of different manifestations in response to interacting with any number of other causal powers of objects that serve as potential reciprocal manifestation partners (Martin, 2007, pp. 54–56). So, for instance, the roundness of a ball is directed at rolling on a surface that has the properties of smoothness and the appropriate density. But the roundness is also directed at making an impression on surfaces with the appropriate elasticity. The roundness is directed at other manifestations when partnered with other properties of objects.

Second, most one-way causal powers are *pleiotropic* because, in causal processes, they contribute to the causal production of many different outcomes (Molnar, 2003, p. 194). For instance, the roundness of a ball not only contributes to the ball's rolling on a smooth, dense surface, but, at the same time, it contributes to the production of certain sounds, and perceptual experiences in an onlooker who may be watching it. The various outcomes are all consequences of the interactions of

the various causal powers responsible for the ball's rolling but also other causal powers in the vicinity with which it interacts.

Finally, the outcomes of causal processes involving the manifestations of a collection of one-way causal powers are *polygenic*. So they are either the causal outcome of the myriad manifesting causal powers enabling a substance to produce an outcome or the manifesting powers are themselves the direct cause of the outcomes.[8]

Some one-way powers are not multi-track and some do not even require a partner to serve as a stimulus for their manifestations. Such *merely spontaneous powers* are manifested without any manifestation partner (e.g., some quantity of strontium-90's power to beta-decay).

*Two-way powers* are quite different from one-way powers. For one, most of the prominent contemporary proponents of two-way powers take them to be emergent powers (see Lowe, 2008, chapter 5; Mayr, 2011, chapter 9; Steward, 2012, chapter 8).[9] More specifically, they are *strongly* emergent and not merely *weakly* emergent. Regarding weak emergence, on the most common account, the *A*s are weakly emergent from the *B*s if and only if a description or explanation of the *A*s cannot be deduced, calculated, computed, or predicted solely from what is known about the *B*s (Heil, 2017, p. 44; Kim, 2010b, pp. 86–87).[10] This does not preclude the possibility of the *A*s being *inductively predictable*. If it has been observed that the *A*s are present when the *B*s are present and we can reliably predict that something that has the *B*s will have the *A*s, the *A*s may still be weakly emergent (Kim, 2010a, p. 13). What is precluded, according to Jaegwon Kim, is *theoretical predictability*. Knowing everything about the *B*s alone is not sufficient for a reliable prediction of the *A*s (Kim, 2010a, p. 14).

For my purposes here, weak emergence is understood as entirely epistemological. If the *A*s are merely weakly emergent from the *B*s, there is no ontological addition with the *A*s. If the *A*s are weakly emergent powers of some complex object whose properties (including the *B*s) are systematically integrated in some way, then I suggest that the *A*s are weakly emergent *derivative* powers of the object (more on derivative powers in section 11.4). If the *A*s are predictable from the *B*s and, hence, not weakly emergent, they are merely *resultant* powers of the complex object.

Strong emergence is ontological. Put in terms of powers, the *A*s are strongly emergent from the *B*s if and only if the *A*s are weakly emergent from the *B*s and contribute some novel powers to their possessors that go beyond those conferred by the *B*s (Heil, 2017, pp. 44–45; Kim, 2010b, p. 87). If the *A*s are strongly emergent, then they are *basic* powers of the object that possesses them (more on this in Section 11.4). While strong emergence implies weak emergence, weak emergence does not imply strong emergence.

The second most important difference between a two-way power and any one-way power is that a two-way power is not directed at any

specific manifestation. A two-way power can be manifested in a different way given the exact same conditions. If a one-way power *P* is directed at a particular manifestation with *P** and is partnered with *P**, then *ceteris paribus P* will always manifest in the same way with *P**. But if a two-way power *Q* is partnered with *Q**, it can be manifested in different ways. So, according to current accounts of two-way powers, if an agent is deciding between *A*-ing and *B*-ing and has reasons that favor both (and even favor one over the other), it is possible that *Q* be manifested in either deciding to *A* or *B*, or even in refraining from doing either. That is to say that the agent possessing *Q* has the power to decide either to *A* or *B* (or refrain from either).

Most contemporary proponents of two-way powers identify two-way powers as *causal* powers that are at least causally *relevant* in the production of action. Some hold that the manifestation of a two-way power causally *enables* an agent *qua* substance to causally produce an outcome such as the acquisition of an intention in decision-making (Mayr, 2011, pp. 213–232). And, along similar lines, others identify actions with the exercise of the two-way power of agency when an agent causes an outcome (Steward, 2012, p. 199).

Others take two-way powers to be non-causal *spontaneous* powers that are responsive to reasons. They are, thus, distinctively *rational powers*. For instance, E.J. Lowe contends that the will is a two-way power. It is a rational power. Exercising this power involves the production of outcomes such as bodily movements (Lowe, 2008, pp. 176–178, 187–190; 2013a, pp. 164–171; 2013b, pp. 173–179). But while this may look like a causal power, it is not, according to Lowe, since a two-way power may be manifested without any effect resulting (2008, p. 150). I follow Lowe in taking a two-way power to be a rational power manifested in response to the presence of reasons for action. Moreover, I expect that most of the current proponents of irreducible two-way powers in theorizing about agency would accept that they are rational powers.[11] But, contra Lowe (2008, pp. 154–157), I contend that even if two-way powers are rational powers, they are *causal* powers whose manifestations do not always have actions as outcomes. In cases of acting, they have mental activity or bodily activity as the outcome of their manifestation. In cases of basic omissions, such powers manifest in an agent's refraining from acting but remain "at the ready" to contribute to causing an output should circumstances change.

Proponents of versions of agent-causalism have been the primary proponents of two-way powers in theorizing about intentional agency. This is even true of Lowe, whose views actually come closer to those of the other proponents of two-way powers than he may have been willing to admit. For instance, he has contended that human agents "can only cause anything *by acting in some* way" (2013b, p. 175). This is very similar to the views of other proponents of two-way powers.[12] For instance,

Erasmus Mayr asserts that "basic physical actions can also be considered as agent-causings" (2011, p. 224). And Helen Steward contends that actions are events "that are causings of bodily movements and changes *by agents*. Causation by actions is therefore just causation by causings by agents" (2012, p. 205).[13] But while recent proponents of two-way powers have almost uniformly endorsed versions of agent-causalism, I cannot think of any reason why a commitment to two-way powers would fail to be compatible with an understanding of them as among the causes that directly *produce* outcomes. While I am not aware of any proponents of such a view, it certainly does not seem ruled out by having a theory of agency that is ontologically committed to the existence of two-way powers.

Finally, while some proponents of two-way powers insist that the manifestation of a two-way power in exercising intentional agency is incompatible with causal determinism (e.g., Lowe 2008; Steward, 2012), others are silent on this question (Alvarez, 2013; Mayr, 2011).[14] For this reason, I will ignore the question of whether or not the exercise of intentional agency is compatible or not with determinism. My target is the putative need for such powers to account for intentional agency, especially in practical decision-making.[15] I will take a neutral stance here on whether or not exercising intentional agency is essentially non-deterministic.

## 11.3 Explanatory and metaphysical worries

Suppose we accept that agents possess ontologically irreducible two-way powers. The relevant type of power, at least in humans, is an irreducible *rational* power that I will assume is at least manifested in practical decision-making when faced with alternatives. *Qua* rational power, it is a power that is manifested in response to reasons for action. And *qua* irreducible power, it is a basic property of an agent.

In this section, I will focus on two difficulties faced by current proposals in the metaphysics of agency that countenance such two-way powers. Specifically, I will first argue that the accounts on offer have difficulty explaining why a two-way power is manifested the way it is in light of an agent's reasons for deciding in a particular way. The second problem is over an ontological commitment of such two-way powers. Specifically, I will argue that a metaphysics of agency that includes irreducible two-way powers is ontologically committed to some species of substance dualism.

### *11.3.1 How do reasons explain the manifestations of two-way powers?*

Proponents of two-way powers say very little about how to explain the manifestations of two-way powers *qua rational* powers. While some

(e.g., Lowe, 2008; Mayr, 2011) say more than others, the accounts they offer leave much to be desired. What is revealing about these accounts is how little they have in common with the actual views of Aristotle. In fact, they are more like the views of Medieval voluntarists. Anyone familiar with the Medieval debate between the intellectualists and the voluntarists will recognize that, if I am right, many prominent proponents of two-way powers today endorse a view that is more voluntarist than Aristotelian. As such, the contemporary theories share the same liabilities of their Medieval antecedents.

While they are understood as spontaneous/active, two-way powers, *qua* rational powers, are not meant by their proponents to be regarded as arational, manifesting indiscriminately. For instance, Lowe clarifies that by characterizing a two-way power as spontaneous what he means is that no prior occurrences can count as the cause of its activation (2008, p. 150). They are, effectively, self-activated. Mayr writes that as an active power, the power of agency is manifested largely independently of "specific external circumstance" (2011, p. 219). My worry here is that, given their overemphasis upon the spontaneity of two-way powers and the independence of their manifestations from the activity of other powers – particularly those that are the constitutive powers of reasons, proponents do not have an adequate story to tell about the manifestation conditions for two-way powers *qua* rational powers. Hence, their theory of intentional agency is explanatorily thin, leaving us with gaps in our understanding of the etiology of intentional agency.

Reasons, on the view of proponents of two-way powers, are *necessary conditions* for the exercise or manifestation of a two-way power (Frost, 2013, p. 613). Differently stated, reasons for action are the *sine qua non* for the exercise of two-way powers *qua* rational powers. Examples of this include Mayr's account on which an agent is following "a standard of success or correctness provided by [their] reason" (2011, p. 292). A standard of success is what the agent follows in their exercise of agency in order to achieve their goal. Lowe's account amounts to a story about the reasons of which an agent was aware at the time they made a decision (2008, pp. 189–190). Lowe balks at the demands of philosophers for a deeper story, comparing those who would not be satisfied with a *just-so* story to "little children" who "sometimes don't know when to stop asking 'why?'" (Lowe, 2008, p. 190). Similarly, Helen Steward writes of reasons as playing an influencing role, with *agents* determining what they will do in response to reasons of which they are aware (2012, p. 151). Things differ little with other proponents of two-way powers. The reasons (whether understood as states of affairs external to the agent that favor a course of action[16] or as internal representational states of an agent[17]) – more specifically, their constituent causal powers – are not active partners that interact with a two-way power in a causal process to generate a particular outcome. So how do they influence the

activation of an agent's two-way power for willing or exercising agency? What is it about reasons in virtue of which a two-way power responds to them if the story of their manifestation is not a causal story like that of one-way powers?

We can contrast the views of contemporary proponents of two-way powers with the position of Aristotle.[18] Aristotle takes states of affairs under which a two-way power may be activated to be merely necessary conditions for its manifestation (*Metaphysics* 1048a, 1–10). As such, a two-way power is conditional because, as David Charles notes, "the agent, constituted as he is, cannot exercise [a two-way power] except in certain circumstances" (1984, p. 24, n. 14; cf. *Metaphysics* 1048a, 15–24). In circumstances under which a two-way power can be manifested, a desire is what will activate the power one way or another (Frost, 2013; Charles, 1984, p. 58). Importantly, the action desired is discerned as good and pursuit-worthy and hence desired under the guise of the good (see *De Anima* 433a, 27–29). The desire is for an action that will allow one to achieve an end set by an agent's thought, including their practical reasoning (*Nicomachean Ethics* 1139a, 31–33). This is a very different account from what we find in the work of today's proponents of irreducible two-way powers. There are no explanatory gaps on this account. The activation of a rational power is part of a larger goal-oriented causal process involving the activity of various states of an agent that together ultimately lead to an exercise of agency.

The proponents of two-way powers on whose work I am focusing here have views that more closely align to those of the Medieval voluntarists than the views of Aristotle. Voluntarism stands in contrast to intellectualism.[19] Intellectualists (e.g., Siger of Brabant and Godfrey of Fontaines) emphasized the primacy of the intellect over the will. The will is moved by the intellect. For instance, while allowing that the will is a rational power that "extend[s] to opposites" (*Quodlibet* 8, q. 6, n. 27),[20] Godfrey of Fontaines argued that if one asserts that "to be able to determine itself belongs to the will from its own nature and not to the intellect, the proposed view is arrived at without a rational basis" (*Quodlibet* 8, q. 6, n. 38). He goes further, arguing that "just as the object of the will actuates the will in line with the way it is apprehended by reason under that aspect under which it is naturally suited to move the will, so too the object of the will determines the will, and the will does not [determine] itself" (*Quodlibet* 8, q. 6, n. 44). Thus, Godfrey seems to take the interaction of the intellect and will (which are both powers[21]) to suffice for the will to be moved (*Quodlibet* 8, q. 6, n. 44; *Quodlibet* 15, q. 4, n. 4).

Voluntarists reverse the order of priority with respect to the relationship between the will and the intellect. They are united in understanding the will as an essentially free power that is not moved by the intellect. William of Auvergne articulates this commitment clearly when he writes that "the will is in itself most free and in its own power in every respect

with regard to its operation …, and for this reason, it is able to correct and direct itself" (2000, p. 97). He goes further, describing the will as like a king with the power to command. Reason is like a counselor offering advice that can be followed or ignored by the will (2000, p. 126). In *Quodlibet* XIV, Question 5, Henry of Ghent defines the will as freer than the intellect. Unlike the freedom of the intellect, the freedom of the will "is the faculty by which it is able to proceed to its act by which it acquires its good from a spontaneous principle in itself and without any impulse or interference from anything else" (Henry of Ghent, 1993, p. 81). Finally, John Duns Scotus – perhaps the most well-known of the voluntarists – in his *Questions on the Metaphysics IX*, q. 15, describes the intellect as "showing and directing" and the will as "inclining and commanding." He goes on to elaborate on their nature, noting that intellect is a one-way power and will a two-way power (1997, p. 141). And in his *Opus oxoniense* II, dist. 42, qq. 1-4; nn. 10-11, Duns Scotus affirms the primacy of the will, writing that "if the will turns towards the same thing as the intellect, it confirms the intellect in its action." He adds in the same text that "the will with respect to the intellect is the superior agent…." (1997, p. 151). Arguably, what we find in the works of the voluntarists is a view similar to Aristotle's in one respect. The intellect provides the circumstances that are the *sine qua non* for the activation of the will as a rational power. But, unlike Aristotle, there is no role for desire as an efficient cause of its activation. Finally, there is no other story that is offered about the activation of the will (such as the one we find in the intellectualists like Godfrey of Fontaines). Why it manifests in one particular way is never explained (apart from a possible *ex post facto* story an agent may tell us that may or may not be accurate).

The similarities between the voluntarist position and what we find in contemporary defenses of two-way powers should not be lost on the reader. The importance of this comparison is that the accounts of two-way powers on offer today have resurrected (whether wittingly or not) an understanding of the will/power of agency on which its activation is left underexplained by an agent's reasons for action. They are underexplained because an agent's reasons, like many other structuring factors that an agent's exercise of a two-way power depends upon, do not provide anything like an explanation for why the power was exercised the way it was.

I should be clear that what I am not demanding is that we have an account of *contrastive* reason-explanations. I am merely stating that we have no account on offer that can allow us to give a principled explanation in terms of an agent's reasons for *why* an agent's power was exercised the way it was and when it was when it was manifested. Neither the accounts offered by Aristotle nor Godfrey of Fontaines face similar difficulties; and, as I hope will be evident by the conclusion of section 11.4 of this chapter, the theory I put forth will not face this worry.

### *11.3.2 A metaphysical worry: The unavoidability of substance dualism for the proponent of two-way powers*

Suppose we accept ontologically irreducible two-way powers of agents in our metaphysic of agency. The relevant type of power is an irreducible rational power. Such rational powers are intrinsic psychological properties. Moreover, such properties, assuming they are basic powers (i.e., fundamental intrinsic dispositional properties of objects), are not ontologically reducible to physical powers. They are strongly emergent powers that are in place when certain conditions are satisfied at the level of the emergence base. If this is right, then a commitment to irreducible two-way powers in our metaphysics of intentional agency implies an ontological commitment to property dualism in our metaphysics of mind, more generally.

Some may regard an ontological commitment to property dualism to be an acceptable consequence since many philosophers these days embrace property dualism under the moniker of "non-reductive physicalism."[22] Many seem to think that property dualism is consistent with substance physicalism. Thus, some of the properties of physical substances are assumed to be mental properties that are not reducible to physical properties. But this view is not as innocent as many seem to think.

Apart from the common worries for property dualism raised by versions of the causal exclusion argument in the literature on mental causation (Kim, 2005), there may be another worry about the ontological implications of property dualism. Specifically, Susan Schneider (2012) has recently argued that property dualism entails substance dualism.[23]

Schneider's argument does not rest so much on whether properties are particulars (tropes or modes) or immanent universals. The problem will be the same regardless of whether properties are understood as particulars or immanent universals.

There are two versions of Schneider's argument. The first assumes a bundle theory of substance on which substances are bundles of properties or property-instances that stand in a relation of compresence to one another. The properties of a substance on such a view are components of the substance. Any addition of properties would affect the identity of any substance. Now, if we accept a dualism of properties, Schneider argues, then, assuming the mind is a substance, the mind is not a physical substance. It is either a hybrid substance or else there are two interacting substances: a physical substance and a mental substance (Schneider, 2012, pp. 64–67). In either case, we have a dualism of substances, not just properties: hybrid substances or mental substances and physical substances.

What if we have irreducible substances as a fundamental ontological category – as substrata, if you will – along with properties? The problem

is the same for a substratum theory of substance as it is for a bundle theory. If intrinsic properties are attributes that characterize objects, then a way that the substances are that have them is non-physical. So if minds are objects that have non-physical intrinsic properties, then they are either hybrid substances or else they are distinctively mental substances that interact with physical substances.

One may think that this problem is avoidable if we understand (*per impossibile*) mental properties as properties of complex physical properties (i.e., collections of functionally integrated physical properties). In that case, a way that the physical properties are would be non-physical. If the oddness of non-physical physical properties is not strange enough, we would still have substance dualism. The reason for this is that one of the ways in which non-physical physical properties would be characterizing the substance that possesses them is as non-physical. Therefore, it seems that an unavoidable consequence of property dualism is some species of substance dualism.

Extending the argument, if ontologically irreducible two-way powers are non-physical properties, which they would be according to their proponents, then a theory of intentional agency that includes irreducible two-way powers entails substance dualism. And emergence will not help here. Emergent powers are basic properties of an object, the agent; and the agent would appear to be an immaterial substance that is not reducible to a physical system of some sort. Thus, a commitment to emergent two-way powers will entail a commitment to emergent substances as well. The only way to avoid this conclusion is to deny that two-way powers are basic powers. But this would amount to accepting an ontologically reductive account of two-way powers, which is precisely what the proponents of irreducible two-way powers wish to avoid.

Admittedly, the argument I have offered is fairly quick. It may be argued that some absurd consequences follow from the conclusion of my argument. Specifically, it may be argued that I have committed the composition fallacy. For instance, if properties are either components of objects or ways that objects are and a property is imperceptible, then it looks like the object that possesses the property is imperceptible.[24] So even if minds are wholly physical systems, a complex object like a mind is imperceptible given that some properties of minds are imperceptible. This looks like a ridiculous consequence of the sort of reasoning I have offered from non-physical properties to substance dualism.

This worry is a chimera, however. While the reasoning above would be an example of the composition fallacy, no such fallacy is committed in my own reasoning. Rather, what I am suggesting is something like the following. If a property is an attribute of an object, a way that it is, then, if an object has non-physical properties, then being non-physical is *a* way that it is. It may be both physical and non-physical. But even if that is so,

then we have a species of substance dualism on our hands – there would be physical substances and hybrid substances. But, it may be argued, this still does not show that we do not face an absurd situation like the one where an imperceptible property possessed by an ordinary human agent would render the agent imperceptible. I think there are good reasons to think this counterargument fails.

Imagine a red ball that weights 1 kilogram. The ball's mass, shape, and color are all intrinsic properties of the ball. They are attributes of the object. That is, they are ways it is. The ball, *qua* substance, is round. Its roundness is not somehow all there is to the ball. Its roundness is not its mass. Its mass is not its color. And so on. So also, in saying that minds are either hybrid or entirely non-physical substances owing to their having non-physical properties is not to say that that is all there is to minds. Rather, the broad types of properties, physical and/or non-physical, can be further divided up into properties that play various functional roles in the system owing to their causal profile. Notice that some of these properties are imperceptible. Returning to the case of the ball, its mass is not perceivable. Rather, we detect the effects of its mass when the ball is weighed (with the mass, the rate of acceleration of gravity on Earth, and the restoring force of the spring all being manifested and serving as the polygenic cause or polygenic enablers of the ball as a cause to register a certain weight when placed on a scale). Now, is the imperceptibility of the ball's mass a property of the ball? It would seem it is not. Rather, it is simply a truth about the mass of any object, that it cannot be perceived. The true modal claim about the imperceptibility of mass need not include among its truthmakers a property of imperceptibility. Any such move betrays a serious mistake about what sorts of ontological claims we are licensed to make on the basis of our true representations. Let me explain.

The mistake made by my interlocutor rests on an all-too-common error in metaphysics of mind. Specifically, the objection seems to assume that if a property *P* is imperceptible, there is some property of imperceptibility that is possessed by a mind that bears *P*. But the mistake being made here is to move from the fact that *P* can be truthfully described as imperceptible to then ascribe a property of imperceptibility to the bearer of *P* (or, worse still, to treating the property *P* as having a property of imperceptibility). That is, my interlocutor is assuming that we can read the properties of an object off what we can truthfully predicate to the object in our talk about the object.

C.B. Martin called this tendency to move from our predicates to properties the error of "linguisticism" and wrote that while it is "silly," it is "also endemic and largely unnoticed by many practising ontologists" (2008, p. 80).[25] He notes that the tendency to linguisticism derives from the uncritical acceptance of a particular understanding of Quine's criterion of ontological commitment:

> [A] theory is committed to those and only those entities to which the bound variables of the theory must be capable of referring in order that the affirmations made in the theory be true.
>
> (Quine, 1948, p. 33)

But it is not clear that this criterion is correct or even very useful for determining what there is. I suggest that instead of Quine's criterion we accept a truthmaker criterion such as the one presented by Ross Cameron according to which "the ontological commitments of a theory are just those things that must exist to *make true* the sentences of that theory" (2008, p. 4). So on the truthmaker criterion, "<*x* exists> might be made true by something other than *x*, and hence that '*a* exists' might be true according to some theory without being an ontological commitment of that theory" (Cameron, 2008, p. 4).

What do debates over the proper criterion of ontological commitment have to do with my interlocutor's objection? '*P* is imperceptible' can be true, but that does not commit us to a *property* of *imperceptibility* possessed by the object that possesses *P* (or, for that matter, a property of imperceptibility possessed by the property *P*). Rather, it is true that '*P* is imperceptible' owing to intrinsic properties of normal perceivers being such that normal perceivers lack the power required to perceive *P*. Notice that this is owing to the kind of property *P* is and the properties of cognitive systems like ourselves. There is no reason to think that *imperceptibility* is an actual component of the object or a *way* that the object in question is. That it is imperceptible is made true by the intrinsic properties of the object and the intrinsic properties of the human visual system.

Before moving to the next section of this chapter, I should be clear that my main goal in the present sub-section has not been to argue against dualism, *per se*. Rather, my goal is simply to bring to the attention of readers an ontological cost of accepting irreducible two-way powers.

In the next section, I present a reductive alternative that takes two-way powers to be derivative properties. The account is free of any obvious ontological commitments to a particular metaphysic of mind. The theory leaves it open for there to be other reasons for accepting property dualism and, hence, substance dualism, or to accept an entirely different metaphysic of mind. For this reason, assuming it can deliver what some philosophers want from two-way powers, it should be preferred over the standard accounts that take two-way powers to be irreducible basic powers of substances.

## 11.4 Reducing two-way powers

In this section, I will construct an alternative, ontologically reductive account of two-way powers. My alternative denies that two-way powers are strongly emergent basic powers. They may be weakly emergent, but

there is no addition of being when an agent can be truthfully described as possessing a two-way power.

The reductive account I will offer does not take two-way powers to be simple conjunctions of one-way powers. Rather, I suggest understanding two-way powers as constellations of what George Molnar refers to as "derivative powers" (2003, p. 143). More specifically, they are *derivative collective powers* of sophisticated agents, which are complex cognitive systems.[26]

Molnar offers the following initial parsing of what he means to pick out by the locution, 'derivative power':

> A power is derivative if the presence of this power in the object depends on the powers that its constituents have and the special relations in which the constituents stand to each other.
>
> (Molnar, 2003, p. 145)

Non-derivative powers are basic powers. Molnar refers to the properties of the simple constituent objects of complex objects as "homogeneous properties." Properties of the complex object as a whole he refers to as "collective properties" (2003, p. 143). If complex objects have basic powers, those powers are powers of emergent substances. Finally, simple objects can have basic and derivative properties. Any derivative powers of simple objects are derived laterally from their other intrinsic properties (Molnar, 2003, p. 143).

Any derivative power is identical with the systematically integrated conglomeration of powers and their relations to one another that together are sufficient for us to ascribe a power to a complex object (Molnar, 2003, p. 144). The derivative power is individuated by its macro-level functional role in the system that is identical with the complex object. Thus, the collection of powers together plays a functional role in a system that provides the truthmakers for describing a complex object as having a functional property that is a derivative power. "The intentional object of [a complex object's] derivative collective power is the same as the intentional object of the jointly exercised powers of the parts of [the complex object] that stand in the relevant special relations" (Molnar, 2003, p. 145). Thus, we get something like the following schema for 'derivative power':

> [An object S's] power to $\phi$ is derivative if the (actual or possible) joint exercise of several powers of some of [S]'s parts, when these parts stand in special relations, manifests $\phi$-ing.
>
> (Molnar, 2003, p. 145)

So, for instance, a molecule has the power to bond owing to at least some of the powers of its constituent atoms and the relations (including existing bonds) they stand in with respect to one another. The power

of a quantity of sodium bicarbonate to neutralize hydrochloric acid is a derivative power it possesses in virtue of the powers of the constituent atoms and the bonds they have with one another. Or consider another example, a human agent's power of visual perception is a psychological property possessed by the agent *qua* animal. The power is possessed in virtue of the powers of their visual system and their relations.

A quick word about derivation and reduction is in order. Molnar suggests that there is a weak sense of 'reduction' that suits all cases of derivation. We have a reduction in this weak sense with all derivative powers when we reduce the number of independent basic powers. Derivative powers are identical with their collective grounds. Two-way powers as derivative powers are reducible to another category of powers, viz., one-way powers.

I suspect that some derivative powers are *weakly* emergent, but they are not *strongly* emergent. So they could not be deduced or (theoretically) predicted solely from information about their basal conditions – specifically, the properties and relations of the microconstituents of the complex object to which the derivative power is attributed. Some derivative powers, however, may be *resultant* powers of a system, being predictable from the microconstituents of a system that is a complex object. Two-way powers, I will assume, are derivative powers that are weakly emergent. From the powers that ground a derivative two-way power, we cannot deduce or (theoretically) predict the two-way power. That said, whether they are actually resultant powers rather than weakly emergent powers will not matter much for my purposes.

Molnar and others are mostly silent on what I will call the *valence* of derivative powers.[27] I assume that basic causal powers of basal simple objects have a valence of 1 on a 0 to 1 scale. Derivative causal powers (such as the constitutive powers of mental states) have a valence anywhere between 0 and 1. I assume that the valence of a derivative causal power is a function of the collection of more basic powers from which it is derived and the causal powers with which it interacts. The strength of a power towards a particular manifestation is, in part, expressible in terms of its valence.

Opposing causal powers may more easily mask a causal power with a low valence. When we have a constellation of partnering powers that are manifesting at a time, a masked power in that process has a valence of 0 owing to its relation with any causal power directed at an opposing manifestation with an equal or greater valence. The subtractive force of a power with an equal valence results in the masking of both interacting powers. In such cases, the two powers are neutralized. But one or more powers with a greater valence than an opposing power that is masked will have a reduced valence owing to the subtractive effect of the interaction with the power that is masked. In effect, the power or powers that serve to mask a power will be partially masked. Finally, some powers have their valence amplified by the additive presence of other powers.

So how might derivative powers allow us to account for our intentional agency? Specifically, how do they provide us with the truthmakers for talk about two-way powers manifesting in decision-making?

Consider the following scenario. Suppose that Soren is craving cherry pie but also wants to avoid needless calories. He is presented with some cherry pie by a host at a dinner party. Ergo, he believes that some cherry pie is available and he believes that to eat the pie will result in his consuming empty calories. I should add that Soren is not on a diet, so he lacks any standing intention the content of which represents a personal policy to avoid eating dessert. Owing to his wanting to make up his mind and his knowledge that circumstances require that he do so, Soren has acquired a proximal intention to decide whether to have pie or refrain from having pie. Thus, in this scenario, we have an agent with *at least* the following mental states: (a) a proximal intention to make up his mind about whether to have pie, (b) a desire to eat some pie and a desire to refrain from consuming empty calories,[28] (c) a belief that both options are presented to him, and (d) a belief that he can either eat the pie or refrain from eating it, but not both.

Soren can be truthfully described as having the two-way power to decide either to eat the pie or abstain from doing so (see Figure 11.1). I will assume that the valence of the two-way power is .5. (But if Soren were more inclined towards one outcome than another, that would be perfectly consistent with the views of some proponents of two-way powers (see, for instance, Steward, 2009; 2012).)

As I have described things so far, Soren is literally *ambivalent*. We have a case where there are the truthmakers sufficient to truthfully describe Soren as possessing a two-way power to decide to either eat the pie or refrain from eating it. Whichever way he decides, he has the

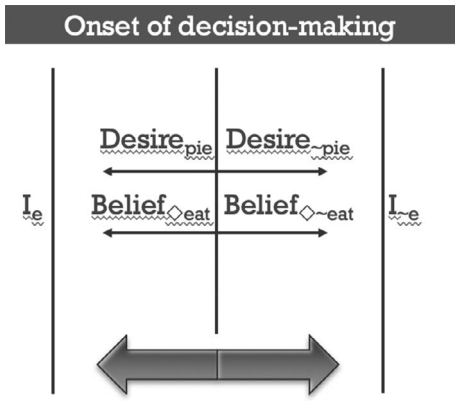


*Figure 11.1* At the onset of decision-making, the overall valence of the two-way power does not favor either of the two possible courses of action.

power to decide differently. And the possession of this power provides the truthmaker for the claim that the alternative decision is possible.[29]

Once a decision is made, we can truthfully say that the two-way power is manifested when Soren either forms an intention to eat the pie or forms an intention to refrain from eating the pie. That this power is manifested is made true by the mutual manifestation of the reciprocal causal powers constitutive of Soren's derivative two-way power to decide to eat the pie or abstain from so doing.

Importantly, in finally acquiring an intention, other powers beyond the constituent powers of the mental states would have also been constitutive of the total constellation of reciprocal powers and it is the various aggregate manifesting powers that are the polygenic cause or causal enablers of the outcome that is the acquisition of the intention (the entire process from Soren's being ambivalent to acquiring the intention would be the mental action of forming an intention). Importantly, these other powers beyond the constituent powers of the mental states contribute to the valence of the interacting powers and the valence of the final outcome (see Figure 11.2). They may tip the balance toward either of the two outcomes. So, for instance, if the pie is especially aromatic, visually appealing, and Soren feels he would not be gluttonizing if he consumes the pie, then the total balance of powers would favor intending to eat some pie. That is, the power to decide to eat the pie would have a greater valence than the opposing power. But given the opposing powers, things could have gone differently. The power to decide not to eat the pie would have a greater valence if the pie were old, visually unappealing, and Soren felt that he could not eat another bite. Finally, if

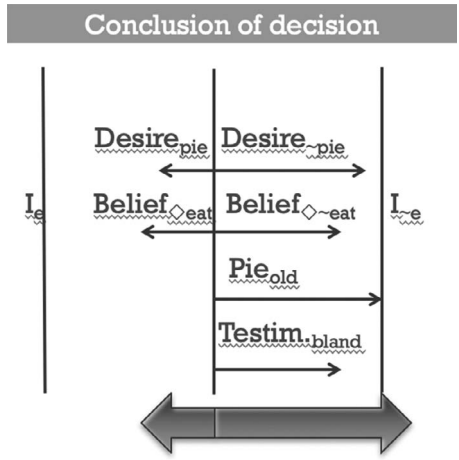


*Figure 11.2* Decision-making concludes when the overall valence of the two-way power favors one course of action over the alternative.

the balance of basic powers were such that neither outcome was favored owing to the additive and subtractive effects of the manifestations of the various causal powers combining with the powers that are constitutive of the opposing motivating states, then we would have a zero-sum outcome. Soren would simply fail to form an intention either way. We would have a non-intentional omission to decide.[30] It would be non-intentional since it would not be the case that Soren omits inadvertently or accidentally. He would knowingly omit to decide but does not intentionally omit to decide.

Why should we favor this account over the standard account of irreducible two-way powers? First, it is less ontologically costly. We get a theory of two-way powers that gives us what we need to truthfully talk about two-way powers without any *ontological commitment* to substance dualism.

Second, the account gives us the tools we need to understand how an agent settles whether or not he decides to *A*. It is the activity of the total agglomeration of basic objects with their powers that make up the functionally integrated system that is the agent that either produces an action or results in an omission from acting. It is up to the agent *qua* functionally integrated system to do what they do. Importantly, while it is up to an agent what they do, it is not a mystery why they decide as they do. We can point to the valence of the causal powers – including the constituent powers of the agent's reasons – that are active in the process of their exercising intentional agency. The account that emphasizes irreducible two-way powers has a disadvantage here. We merely have a "just-so" story given by some prominent defenders about why the agent decides in any given way and there is no deep story about the interaction between the agent's two-way power with the other powers constitutive of the agent. Agency on such a view looks more like an arational process involving a merely spontaneous power like the alpha decay of some quantity of uranium-238.

Finally, some may worry that this is a deterministic theory. But what I have offered here is a reductive account of two-way powers in agency that is consistent with a causal powers metaphysic of causation on which causal relations are non-necessitating (see Mumford & Anjum, 2011).[31] So while the account I have presented may be consistent with some compatibilist intuitions about intentional agency and determinism, nothing about the account rules out an incompatibilist understanding of intentional agency.

## 11.5 Two objections

I am certain that proponents of irreducible two-way powers will not find the account on offer attractive. In this section, I consider only two objections. I do not think that either proves fatal to my theory of two-way powers.

### *11.5.1 Conjunctions of powers and constellations of powers*

It may be argued that my reductive alternative is for all intents and purposes just a variant of the conjunctive strategy on which a two-way power is just a conjunction of one-way powers. I maintain that my strategy is *not* a version of the conjunctive strategy since, on my account, the two-way power is a *constellation* of multiple reciprocal powers with a total valence that is a function of the interacting constituent powers of the derivative power

Suppose, however, that my account could reasonably be interpreted as a conjunctive theory of two-way powers. This would be so given that the constellation of powers with which a two-way power is identical includes some powers that favor *A*-ing and some that favor not-*A*-ing. It may be argued that the following worry raised by Maria Alvarez will prove fatal to the account on offer in this paper:

> It may be tempting to think that one can understand a two-way power as the conjunction of two one-way powers. But this is not so. For one-way powers are characterized by the fact that when the conditions for their manifestation obtain, the power will be necessarily manifested. But if an agent had the ability and opportunity to $\phi$ and also the ability and opportunity not to $\phi$ at $t$, and this were the conjunction of two one-way powers, then the agent would both $\phi$ and not $\phi$ at $t$ – but that is impossible.
>
> (Alvarez, 2013, p. 109)

This type of reasoning may be compelling to some. But the threat is chimeric.

There is no good reason to think that if the power to decide is a conjunction of powers, then an agent with such a power will *both* decide to *A* and refrain from deciding to *A*. That Eva will decide one way or another depends upon the constellation of reciprocal causal powers constitutive of her reasons and power to decide. She will simply omit to decide if the balance of her total reasons do not favor *A*-ing over not-*A*-ing. And she will decide to *A* if the balance of her total reasons favor *A*-ing. But whether she will decide to *A* or do otherwise (including refraining from deciding either way), will depend upon the total valence of the derivative two-way power toward one or the other outcome. That the agent would do both *A* and not-*A* in cases of ambivalence is clearly not an implication of the view.

### *11.5.2 If there is no objective chance that someone will decide differently than they do, they are not exercising a two-way power*

It may be argued that what I have presented is hardly a replacement for understanding two-way powers as irreducible. Specifically, the worry is

that we have lost anything that can truthfully be described as a two-way power (even if a derivative power) on my account. The worry may best be illustrated by an example.

Suppose that, in the case of Soren, he decides not to have pie. Assume, further, that the combined valence of the interacting powers manifesting in the causal process of his decision-making was close to 1. While Soren still has a desire for pie, etc., the valence of the constituent powers favoring pie is so low that, while it is possible that Soren decide differently, the possibility is remote. If that is the case, it hardly seems like Soren can be truthfully described as having had the power to decide to have the pie or not have the pie.

The worry seems misplaced. If we are actualists and are looking for truthmakers for modal claims in metaphysics, we have them in the powers of objects (including agents). Strictly speaking, Soren has powers that, together, are directed at deciding differently than he actually does. These are sufficient for making it true that, while making a decision, he has the power to decide to have pie or not to have the pie (see Aristotle, *Metaphysics Θ*, 1048a, 10–24).

Finally, it is instructive at this point to note that at least some proponents of two-way powers have made claims even stronger than what I am making. For instance, Helen Steward contends that the possession of a power is not the same thing as there being an objective chance that one will manifest that power. She writes that thinking that "having the power to ϕ requires the existence of some objective chance that one will ϕ [is a mistake] since where what puts one's ϕ-ing quite out of the question is only such things as one's own wants, principles, motivations, etc." (2011, p. 126). If the view I have offered is problematic, this sort of position that is favored by a prominent proponent of robust, irreducible two-way powers will face similar troubles. Therefore, the remoteness of one's deciding differently should not be a problem. There is still the genuine possibility of one deciding differently at the time one decides on my reductive account.

## 11.6 Conclusion

Assuming that we need two-way powers for understanding our agency in decision-making (and our intentional agency, more generally), we have an alternative to the doctrine that two-way powers are basic causal powers that are not reducible to more fundamental powers. On my account, a two-way power is a derivative power that is ontologically reducible to a constellation of reciprocal causal powers. We have, then, two options on the table. Either two-way powers are irreducible, strongly emergent, *sui generis* powers, or else they are derivative powers that are ontologically reducible to constellations of one-way powers. Which one should we accept?

Given the implied ontological commitments of irreducible two-way powers, I maintain that considerations of parsimony and ontological costliness alone should lead us to accept the reductive account. We get to truthfully deploy the concept of two-way powers when we invoke the concept of intentional agency in decision-making without any commitment to either more than one type of power or to a substance dualist metaphysic of mind, and the account does not suffer from the sorts of explanatory gaps we find in non-reductive theories of two-way powers.[32]

## Notes

1. The edition I am here relying on is C.D.C Reeve's 2016 translation.
2. See, for instance, Alvarez (2013), Kenny (1989), Lowe (2008; 2013a; and 2013b), Mayr (2011), Pink (2008), and Steward (2012).
3. Whether any reduction of the concept of a two-way power is actually possible is something I leave as an open question. I have no settled views on this matter. Hence, I shall simply assume that it is an irreducible concept and leave the question of conceptual reduction for another occasion.
4. See Frost (2020) for another recent critique of some current defenses of two-way powers. Specifically, he argues that recent proposals cannot avoid collapsing into accounts involving the activity of one-way powers. He suggests that a theory closer to the view proposed by Aristotle can avoid the challenges he raises.
5. Henceforth in this paper I will largely dispense with referring to 'ontological reduction' and its cognates. Unless specified otherwise, I should be understood to be referring to *ontological reduction* by the term 'reduction.'
6. Most of the authors I discuss either explicitly endorse an account of two-way powers on which their exercise extends to mental actions (e.g., Kenny, 1989; Lowe, 2008; Steward, 2012) or, while they are silent on mental actions (e.g., Mayr, 2011), their theory can be extended to mental agency. Oddly, at least one – viz., Maria Alvarez (2013) – apparently denies that mental actions involve the activity of two-way powers. This is especially odd since Alvarez identifies exercising agency with exercising such a power. The implication is that mental actions like deciding how to act are not exercises of agency since they do not consist in causing the right sort of event (Alvarez, 2013, pp. 106–107). So while Alvarez will continue to be mentioned, I do so being mindful of the fact that she would resist agreeing with me that practical decisions are good candidates to focus on in thinking about two-way powers.
7. With Martin (2008), Heil (2002; 2012), Jacobs (2011), Jaworski (2016), Mumford (1998), and others, I assume that all of the properties of objects are not just dispositional properties, but they are *powerful qualities*. So there are no purely dispositional properties or purely categorical properties. Rather, under one description all properties are categorical and under another they are dispositional. But this assumption is not important for what I am doing here. (And for an argument to the effect that the distinction between pure powers and powerful qualities is not ontologically deep, see Taylor (2018).) For my purposes, I just merely need the assumption that some properties of objects are dispositional properties and that the causal powers of objects are identical with such properties.

8. See Ingthorsson (2002), Molnar (2003), Chakravarrty (2005), Mumford and Anjum (2011), Williams (2014), Heil (2012), and Buckareff (2017) for accounts of causation that focus on the interacting causal powers of objects manifesting and cooperating to produce polygenic outcomes. See Ingthorsson (2021), Kuykendall (2019), Lowe (2008) and Whittle (2016) for accounts of causal powers enabling a substance to cause outcomes. See Buckareff (2017) for a critique of Whittle (2016). See Kuykendall (2019) for a response to Buckareff (2017).
9. In conversation on November 17, 2017, at the Mental Action and Metaphysics of Mind workshop at the University of London, Maria Alvarez expressed a commitment to taking two-way powers as emergent powers.
10. Mark Bedau offers a slightly more robust account of weak emergence. The *A*s would be weakly emergent from the *B*s if and only if the *A*s can be derived o*nly by simulation* from the *B*s and the external conditions of the system of which the *B*s are a part (Bedau, 1997, p. 378).
11. The only possible exception is Helen Steward (2012).
12. And all of these authors approximate the position of Thomas Reid, who advocated two-way powers (see 1788/1969, p. 259). Reid identified what he referred to as "active power" as "a quality in the cause which enables it to produce the *effect*." And he identified the manifestation of active power in producing effects with "*action, agency, efficiency*" (1788/1969, p. 268).
13. The reference to bodily movements is deceptive since it may suggest to some that Steward does not think that mental actions are caused this way. But Steward is clear that she takes mental actions to involve a kind of bodily movement in Steward (2012, pp. 32–33).
14. For a defense of an explicitly compatibilist account of two-way powers in agency, see Frost (2013). In correspondence (June 30, 2017), Frost indicated that he is non-committal about the existence of irreducible two-way powers but has doubts.
15. Some who favor a view of agency based on an ontology of causal powers that eschews two-way powers argue that we should understand causal processes in agency to be nondeterministic. See Mumford and Anjum (2014; 2015a; 2015b). For replies to Mumford and Anjum, see Franklin (2014) and Mackie (2014). Mumford and Anjum (2011) take all causal production to be nonnecessitating. For critique, see Williams (2014). For a proposal that shares common features with Mumford and Anjum's view but is still quite different (emphasizing a causal role for the reasons of agents as causally structuring the propensity of agent-causal power to produce an outcome with an objective probability between (0,1)), see O'Connor (2009a; 2009b) and Jacobs and O'Connor (2013). Others who endorse a causal powers theory of causation in their theory of agency are silent on whether causal processes involving manifesting causal powers involve the necessitation of outcomes. See, for instance, Buckareff (2018) and Stout (2002; 2006; 2007; 2010; 2012).
16. This is the view assumed by Alvarez (2010), Lowe (2008), and Mayr (2011).
17. Steward (2012) at least allows that reasons may be the internal representational states of agents. Kenny (1989) treats wants of agents as reason-giving.
18. For further discussion of Aristotle's views and how they compare to those of Helen Steward, see Frost (2013).
19. For an excellent brief survey of the Medieval debate, see Hoffman (2010). Of course, as with any such debate in the history of philosophy, there are some figures who resist easy classification as either voluntarists or intellectualists; and the views of some vary in different sources.

20. All of the translations of Godfrey of Fontaines that follow are from Neil Lewis's translation (2019).
21. The intellect is a "free power in reason" (*Quodlibet* 8, q. 6, n. 111). The will is a rational power that is responsive to reasons given by the intellect (*Quodlibet* 8, q. 6, n. 27).
22. Jaegwon Kim has contended that the rise of non-reductive physicalism as the "new orthodoxy simply amount[s] to the resurgence of emergentism" (2010b, p. 10).
23. The only defender of two-way powers of whom I am aware who has defended a version of (non-Cartesian) substance dualism and was sensitive to the connection between a commitment to dualism about properties and dualism about substances is E.J. Lowe (2006; 2008).
24. Matthew Boyle raised this objection at the conference on Mental Action and the Ontology of Mind Conference at the University of London at which this paper was first presented.
25. Heather Dyke (2007) calls this "the representational fallacy" and John Heil (2003) has referred to it as "the picture view" of reality.
26. See Aguilar and Buckareff (2016) for a defense of a gradualist metaphysic of agents that allows for agency to be scalar and admitting of various degrees of sophistication.
27. What is denoted by "valence" in chemistry is similar to what I have in mind here. This locution is not commonly deployed in the literature on the ontology of causal powers and the general phenomenon it describes is often glossed over. The closest thing I have found to what I am referring to as the valence of a power is Harré and Madden's brief discussion of "augmented" and "diminished" powers (1975, p. 95).
28. I recognize that some have pointed to a functional distinction between wants and desires. Moreover, "desire" can have multiple senses. For instance, Wayne Davis (1986) distinguishes "appetitive desires" from "volitive desires." But, for my purposes, I am using "want" and "desire" interchangeably to denote what Davis refers to as "volitive" desire. Volitive desire is a mental state with a world-to-mind direction of fit and with a functional role that is manifested in the acquisition of intentions. Of course, there is a close connection between appetitive and volitive desire (consider the case of Soren and the pie).
29. See Borghini and Williams (2008) for more on powers as truthmakers for modal claims.
30. For discussion of omissions and causal processes in exercising agency, see Buckareff (2018).
31. See footnote 15 above for references on the debate over whether a causal powers theory of causation implies an understanding of causation as non-necessitating.
32. Earlier versions of this paper were presented at the Conference on Mental Action and the Ontology of Mind at the University of London in 2017, in philosophy colloquia at Bard College and Vassar College in 2018, and at the Third Workshop on Agency in the Mountains at Copper Mountain, Colorado in 2019. I am especially grateful to my commentator at the London conference, Lillian O'Brien, for her incisive comments on my paper. I am also grateful to the members of the audience on each occasion that I presented on this paper for their feedback. On the occasions I presented this paper I received very helpful feedback (in the form of objections raised) from Maria Alvarez, Matthew Boyle, Jay Elliott, Alex Grzankowski, Jeffrey Seidman, and Sebastian Watzl. I am very grateful to these audience members for their remarks. Finally, I wish to thank Kim Frost for his very helpful written comments on a draft of this paper.

## References


Aguilar, J., & Buckareff, A. (2015). A gradualist metaphysics of agency. In *A.* In C. Buckareff, S. Moya, & Rosell (Eds.), 30 (pp. 30–43). New York: Palgrave-Macmillan.

Alvarez, M. (2013). Agency and two-way powers. *Proceedings of the Aristotelian Society 113*, 101–121.

Aristotle (2016). *Metaphysics*. Trans. C.D.C. Reeve. Indianapolis: Hackett.

Aristotle (1984). *On the soul* [*De Anima*]. J.A. Smith (trans.). In J. Barnes (Ed.). *the complete works of Aristotle: The revised Oxford translation*, volume one (pp. 641–692). Princeton, NJ: Princeton University Press.

Aristotle (1984). *Nicomachean ethics*. W.D. Ross and J.O. Urmson (trans.). In J. Barnes (Ed.). *the complete works of Aristotle: The revised Oxford translation,* volume one (pp. 1729–1867). Princeton, NJ: Princeton University Press.

Bedau, M. (1997). Weak emergence. *Philosophical Perspectives*, *11*, 375–399.

Borghini, A., & Williams, N. (2008). A dispositional theory of possibility. *Dialectica*, *62*, 21–41.

Buckareff, A. (2011). How does agent-causal power work? *The Modern Schoolman (Now Res Philosophica): Special Issue on Free Will and Moral Responsibility*, *88*, 105–21.

Buckareff, A. (2017). A critique of substance causation. *Philosophia*, *45*, 1019–1026.

Buckareff, A. (2018). I'm just sitting around doing nothing: On exercising intentional agency in omitting to act. *Synthese*, *195*, 4617–4635.

Cameron, R. (2008). Truthmakers and ontological commitment: Or how to deal with complex objects and mathematical ontology without getting into trouble. *Philosophical Studies*, *140*, 1–18.

Chakravartty, A. (2005). Causal realism: Events and processes. *Erkenntnis*, *63*, 7–31.

Charles, D. (1984). *Aristotle's philosophy of action*. London: Duckworth.

Clancy, S. (2013). A strong compatibilist account of settling. *Inquiry*, *56*, 653–665.

Davis, W. (1986). The two senses of desire. In J. Marks (Ed.), *The ways of desire: New essays in philosophical psychology on the concept of wanting* (pp. 63–82). New York: Routledge.

Duns Scotus, J. (1997). *Duns Scotus on the will & morality*. Trans. A. Wolter and W. Frank. Washington, DC: Catholic University of American Press.

Dyke, H. (2008). *Metaphysics and the representational fallacy*. New York: Routledge.

Franklin, C. (2014). Powers, necessity, and determinism. *Thought: A Journal of Philosophy*, *3*, 225–229.

Frost, K. (2013). Action as the exercise of a two-way power. *Inquiry*, *56*, 611–624.

Frost, K. (2020). What could a two-way power be? *Topoi*, *39*, 1141–1153

Godfrey of Fontaines (2019). *Godfrey of Fontaines and the freedom of the will.* Trans. Neil Lewis. http://lewis.georgetown.domains/

Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity*. Oxford: Basil Blackwell.

Heil, J. (2003). *From an ontological point of view*. New York: Oxford University Press.

Heil, J. (2012). *The universe as we find it*. New York: Oxford University Press.

Heil, J. (2017). Downward causation. In M. Paoletti, & F. Orilia (Eds.), *Philosophical and scientific perspectives on downward causation* (pp. 42–53). New York: Routledge.

Henry of Ghent (1993). *Quodlibetal questions on free will*. Trans. R. Teske. Milwaukee, WI: Marquette University Press.

Hoffman, T. (2010). Intellectualism and voluntarism. In R. Pasnau (Ed.), *The Cambridge history of medieval philosophy* (pp. 415–427). New York: Cambridge University Press.

Ingthorsson, R. (2002). Causal production as interaction. *Metaphysica*, *3*, 87–119.

Ingthorsson, R. (2021) *A powerful particulars view of causation*. New York: Routledge.

Jacobs, J. (2011). Powerful qualities, not pure powers. *The Monist*, *94*, 81–102.

Jacobs, J., & O'Connor, T. (2013). Agent-causation in a neo-Aristotelian metaphysics. In S. Gibb, R. Ingthorsson, & E.J. Lowe (Eds.), *Mental causation and ontology* (pp. 173–192). New York: Oxford University Press.

Jaworski, W. (2016). *Structure and the metaphysics of mind: How hylomorphism solves the mind-body problem*. New York: Oxford University Press.

Kenny, A. (1989). *The metaphysics of mind*. New York: Oxford University Press.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.

Kim, J. (2010a). Making sense of emergence. In J. Kim (Ed.), *Essays in the metaphysics of mind* (pp. 8–40). New York: Oxford University Press.

Kim, J. (2010b). "Supervenient and yet not deducible": Is there a coherent concept of ontological emergence. In J. Kim. (Ed.), *Essays in the metaphysics of mind* (pp. 85–104). New York: Oxford University Press.

Kuykendall, D. (2019). Powerful substances because of powerless powers. *Journal of the American Philosophical Association*, *5*, 339–356.

Lowe, E. J. (2006). Non-cartesian substance dualism and the problem of mental causation. *Erkenntnis*, *65*, 5–23.

Lowe, E. J. (2008). *Personal agency: The metaphysics of mind and action*. New York: Oxford University Press.

Lowe, E. J. (2013a). Substance causation, powers, and human agency. In S. Gibb, J. Lowe, & R. Ingthorsson (Eds.), *Mental causation and ontology* (pp. 153–172). New York: Oxford University Press.

Lowe, E. J. (2013b). The will as a rational free power. In R. Groff, & J. Greco (Eds.), *Powers and capacities in philosophy: The new Aristotelianism* (pp. 172–185). New York: Routledge.

Mackie, P. (2014). Mumford and Anjum on incompatibilism, powers, and determinism. *Analysis*, *74*, 593–603.

Martin, C. B. (2008). *The mind in nature*. New York: Oxford University Press.

Mayr, E. (2011). *Understanding human agency*. New York: Oxford University Press.

Mele, A. (2003). *Motivation and agency*. New York: Oxford University Press.

Mele, A. (2017). *Aspects of agency: Decisions, abilities, explanations, and free will*. New York: Oxford University Press.

Molnar, G. (2003). *Powers: A study in metaphysics*. New York: Oxford University Press.

Mumford, S. (1998). *Dispositions*. New York: Oxford University Press.

Mumford, S., & Anjum, R. (2011). *Getting causes from powers*. New York: Oxford University Press.

Mumford, S., & Anjum, R. (2014). A new argument against compatibilism. *Analysis*, *74*, 20–25.

Mumford, S., & Anjum, R. (2015a). Freedom and control: On the modality of free will. *American Philosophical Quarterly*, *52*, 1–11.

Mumford, S., & Anjum, R. (2015b). Powers, non-consent, and freedom. *Philosophy and Phenomenological Research*, *91*, 136–152.

O'Connor, T. (2009a). Agent-causal power. In T. Handfield (Ed.), *Dispositions and causes* (pp. 189–214). New York: Oxford University Press.

O'Connor, T. (2009b). Degrees of freedom. *Philosophical Explorations*, *12*, 119–25.

Pink, T. (2008). Intentions and two models of human action. In B. Veerbek (Ed.), *Reasons and intentions* (pp. 153–179). Aldershot: Ashgate Publishing.

Schneider, S. (2012). Why property dualists must reject substance physicalism. *Philosophical Studies*, *157*, 61–76.

Steward, H. (2009). The truth in compatibilism and the truth of libertarianism. *Philosophical Explorations*, *12*, 167–179.

Steward, H. (2012). *A metaphysics for freedom*. New York: Oxford University Press.

Quine, W. (1948). On what there is. *The Review of Metaphysics*, *2*, 21–38.

Stout, R. (2002). The right structure for a causal theory of action. *Facta Philosophica*, *4*, 11–24.

Stout, R. (2006). *Action*. Durham: Acumen.

Stout, R. (2007). Two ways to understand causality in agency. In A. Leist (Ed.), *Action in context* (pp. 137–153). New York: de Gruyter.

Stout, R. (2010). What are you causing in acting. In J. Aguilar, & A. Buckareff (Eds.), *Causing human actions: New perspectives on the causal theory of action* (pp. 101–113). Cambridge, MA: MIT Press.

Stout, R. (2012). Mechanisms that respond to reasons: An Aristotelian approach to agency. In F. O'Rourke (Ed.), *Human destinies: Philosophical essays in memory of gerald hanratty* (pp. 81–97). Notre Dame: University of Notre Dame Press.

Taylor, H. (2018). Powerful qualities and pure powers. *Philosophical Studies*, *175*, 1423–1440.

William of, A. (2000). *The soul.* Trans. R. Teske. Milwaukee, WI: Marquette University Press.

Williams, N. (2014). Powers: Necessity and neighborhoods. *American Philosophical Quarterly*, *51*, 357–371.

# 12 Are Practical Decisions Mental Actions?

*Alfred R. Mele*

## 12.1 Introduction

Elsewhere, I have developed the idea that decisions about what to do (and what not to do) are momentary mental actions of intention formation, and I have argued that such mental actions exist (Mele, 2000; 2003, ch. 9; 2017, ch. 2).[1] Following Arnold Kaufman (1966, p. 25), I called such decisions *practical* decisions and distinguished them from what he calls *cognitive* decisions: for example, a detective's decision, after investigation and reflection, that a suspect is probably lying about a certain matter. (This point having been made, I feel confident that I can occasionally drop the adjective "practical" and count on readers to supply it.) In arguing for the existence of practical decisions, as I conceive of them, I appealed to ordinary experiences and responded to some action-theoretic worries about the existence of mental actions of intention formation. In this chapter, I explore a worry of another kind, one inspired by various neuroscience experiments.

## 12.2 Background

In earlier work (Mele, 2000; 2003, ch. 9; 2017, ch. 2), I examined four competing views about practical deciding. Here, I discuss just two of them.

*View 1: practical deciding as nonactional.* To decide (not) to *A* is simply to acquire an intention (not) to *A* on the basis of practical reflection (or in some other way), and acquiring an intention – in this way or any other – is never an action. In some spheres, acquiring an *x* may have both an actional and a nonactional mode. Joe's acquiring a car, for example, may or may not be an action of his. In buying my car, he performs an action of car acquisition; if, instead, I give him my car, his acquiring it is not an action of his. The sphere of intention acquisition is not like this; it is one-dimensional and nonactional.[2]

A brief sketch of a familiar nonactional conception of cognitive deciding provides some background for view 1. On this conception, which is

DOI: 10.4324/9780429022579-13

sometimes presented as supporting a corresponding conception of practical deciding, to decide that *p* is the case is simply to acquire a belief that *p* is the case on the basis of reflection (see O'Shaughnessy, 1980, vol. 2, pp. 297–302). In my earlier example, the detective's deciding that the suspect is probably lying is a matter of his acquiring a belief that this is probable on the basis of reflection on considerations that he takes to be relevant. The detective's belief is a product of various actions he performed, but his acquiring that belief is not itself an action. It is not the case that, on the basis of his reflection, he performs an action of belief formation. According to a nonactional view of cognitive deciding, acquiring a belief on the basis of reflection is never an action. View 1 is an analogous view about practical deciding.

*View 2: practical deciding as a momentary mental action of intention formation*. Practical deciding is essentially actional. It is a momentary mental action of intention formation. Practical reflection is not part of any action of deciding, although such reflection may often precede and inform instances of practical deciding. And deciding to *A* does not precede the onset of the intention to *A* formed in the act of deciding. Instead, what it is to decide to *A* is to form – actively – an intention to *A*.[3] The intention arises *in* that momentary intention-forming action, not after it. (The same goes for deciding not to *A*.)

I say a bit more about view 2 here as background for what is to come. Deciding what to do, as I conceive of it, is prompted partly by uncertainty about what to do (Mele, 2000; 2003, ch. 9; 2017, ch. 2).[4] (Being uncertain about what to do should be distinguished from not being certain about what to do. Hurricanes are neither certain nor uncertain about anything.) Setting aside science fiction and the like, when there is no such uncertainty, no decisions will be made. This is not to say that, in such situations, no intentions will be acquired.

Not all intentions are formed in acts of deciding. Consider the following: "When I intentionally unlocked my office door this morning, I intended to unlock it. But since I am in the habit of unlocking my door in the morning and conditions … were normal, nothing called for a *decision* to unlock it" (Mele, 1992, p. 231). If I had heard a disturbance in my office, I might have paused to consider whether to unlock the door or call the authorities instead, and I might have decided to unlock it. But given the routine nature of my conduct, there is no need to posit an action of intention formation in this case. As I see it, my intention to unlock the door was acquired without having been actively formed. That is, it was *nonactionally acquired*.

In attempting to understand practical deciding, one should not be looking for means by which agents form intentions. If there are basic actions – roughly, actions that an agent performs, but not by means of doing something else – momentary actions of intention formation are among them. In Mele (2000), I suggested that a way to try to home

in on practical deciding is to catalog ways in which intentions arguably are nonactionally acquired and to see what conceptual space might remain for the actional acquisition of intentions – that is, for practical deciding. I have identified one item in the catalog already. I turn now to some others.

One might argue that beliefs about what it is best to do that issue from practical reasoning often issue directly in corresponding intentions, and so without any intervening action of intention formation. In some cases, having judged or decided on the basis of practical reflection that it would be best to *A*, one seemingly does not need to proceed to do anything to bring it about that one intends to *A*. The judgment may issue immediately and by default in the intention (Mele, 1992, ch. 12).

Before moving forward in the catalog, I observe that the connection between judgment and intention is not always so smooth. Consider Joe, a smoker (Mele, 2000, pp. 84–85). On New Year's Eve, he is contemplating kicking the habit. Faced with the practical question what to do about his smoking, Joe is deliberating about what it would be best to do about this. It is clear to him that it would be best to quit smoking at some point, but as yet he is unsure whether it would be best to quit soon. Joe is under a lot of stress, and he worries that quitting smoking might drive him over the edge. Eventually, he decides that it would be best to quit – permanently, of course – by midnight. That decision settles an evaluative question. But Joe is not yet settled on quitting. He tells his partner, Jill, that it is now clear to him that it would be best to stop smoking, beginning tonight. She asks, "So is that your New Year's resolution?" Joe sincerely replies, "Not yet; the next hurdle is to *decide* to quit. If I can do that, I'll have a decent chance of kicking the habit."

This story is coherent. In some instances of akratic action, one intends to act as one judges best and then backslides (Mele, 1987, 2012). In others, one does not progress from judging something best to intending to do it.[5] Seemingly, having decided that it would be best to quit smoking by midnight, Joe may or may not form the intention to do so. His actively forming the intention, as opposed to his nonactionally acquiring it, would be a momentary mental action of the kind I call practical deciding.

Here is another item for the catalog. Arguably, some intentions nonactionally arise out of desires (Audi, 1993, p. 64). Ann has just acquired a proximal desire to *A* – a desire to *A* straightaway.[6] Perhaps, if she has no (significant) competing desires and no reservations about *A*-ing, the acquisition of that desire may directly give rise to the nonactional acquisition of a proximal intention to *A*. Walking home from work, Ann notices her favorite brand of beer on display in a store window. The sight of the beer prompts a desire to buy some, and her acquiring that desire issues directly in an intention to buy some (Mele, 2000, p. 86). This is conceivable.

It also is conceivable that, given Ann's psychological profile, the sight of the beer in the window issues *directly* in an intention to buy some, in which case there is no intervening desire to buy the beer (Mele, 2000, p. 86). Perhaps in some emergency situations, too, a perceptual event, given the agent's psychological profile, straightaway prompts an intention to *A*. Seeing a dog dart into the path of his car, an experienced driver who is attending to traffic conditions may immediately acquire an intention to swerve (Mele, 2000, pp. 86–87). This is conceivable too.

I mentioned a connection between practical deciding and uncertainty: as I conceive of the former, it is prompted partly by uncertainty about what to do (except in some very strange cases; see note 4). This connection helps to account for the plausibility of the judgments I entertained about the items in the catalog. In the cases I described, there is no uncertainty that intention acquisition resolves. I was not uncertain about whether to unlock my door, Ann was not uncertain about whether to buy the beer, and the driver was not uncertain about what course of action to take. At no point in time were any of us uncertain about the matters at issue.[7] Furthermore, if there are cases in which a judgment based on practical reflection issues directly (and therefore without the assistance of an act of intention formation) in a corresponding intention, the agent's reaching his judgment resolves his uncertainty about what to do. Reaching the judgment directly results in settledness on a course of action (or, sometimes, in settledness on *not* doing something). In Joe's case, of course, matters are different: even though he has decided that it would be best to quit smoking by midnight, he continues to be unsettled about whether to do that.

I realize that there are those who would deny that there is any interesting connection between deciding and uncertainty. Hungry Henry just now ordered a vegan Cobb salad after looking at his lunch menu. It was the only meal on the menu that he would even consider eating, and he had no interest in looking for another restaurant and no interest in skipping lunch. So he was not at all uncertain about what to do about lunch. Even so, some people would contend that Henry decided to order the Cobb salad. As I see it, they use the word "decide" differently than I do. I have no wish to argue that they should switch to my usage. Instead, I announce that I am limiting the discussion of practical decisions here to those that are responses to uncertainty about what to do.[8]

In previous work on deciding (Mele, 2000; 2003, ch. 9; 2017, ch. 2), I reported on some common experiences of mine that at least seem to be of practical deciding, as I conceive of such deciding. (I hoped that my reports would resonate with many readers.) In an effort to ascertain whether my experiences might be veridical, I examined grounds for skepticism about them. I close this section by rehearsing the reports.

Sometimes I find myself with an odd hour or less at the office between scheduled tasks or at the end of the day. Typically, I briefly reflect on

what to do then. I find that I do not try to ascertain what it would be *best* to do at those times: this is fortunate, because settling that issue might often take much more time than it is worth. Instead, I look at a list that I keep on my desk of short tasks that need to be performed sooner or later – reply to an e-mail message, write a letter of recommendation, and the like – and decide which to do. So, at least, it seems to me. Sometimes I have the experience not only of settling on a specific task or two but also, in the case of two or more tasks, of settling on a particular order of execution.

I have an e-mail system that makes a sound when a message arrives. Occasionally, when I hear that sound, I pause briefly to consider whether to stop what I am doing and check the message. Sometimes I have the experience of deciding to check it, or the experience of deciding not to check it. Sometimes I do not even consider checking the new message.

In situations of both of the kinds mentioned (the odd hour and incoming e-mail), I sometimes have the experience of having an urge to do one thing but deciding to do another instead. For example, when I hear that a new e-mail message has arrived, I may have an urge to check it straightaway but decide to finish what I am doing first. When I am looking at my list of short tasks at the beginning of an odd hour, I may feel more inclined to perform one of the more pleasant tasks on my list but opt for a less pleasant one that is more pressing.

In each of these cases, I experience my deciding to *A* as something I actively do – that is, as an action. Presumably, many readers have similar experiences. But can those experiences be trusted?

## 12.3 Some studies and some questions

It is sometimes claimed that we have good scientific reason to believe that we make all of our practical decisions unconsciously and that, in those cases in which we become aware or conscious of them, we do so after they have been made.[9] Claims about how long the lag time is between decision-making and the onset of consciousness of the decision – or the intention formed therein – range from about a third of a second (Libet, 1985, 2004) to several seconds (Soon, Brass, Heinze, & Haynes, 2008). I have disputed these claims elsewhere, arguing (among other things) that we do not have good evidence for the claim that decisions are made at these early times and that we have better evidence that if pertinent decisions are made, they are made around the time the experimental participants say they make them (Mele, 2009, 2018).[10]

Despite my skepticism about these claims, I believe that the studies at issue raise some interesting questions, including the following one. How reliable are our experiences of decision-making? And, more precisely, how reliable are these experiences, if practical decisions are conceived of as momentary mental actions of intention formation?

A brief description of some well-known experiments by Benjamin Libet (1985, 2004) helps set the stage. Participants are asked to flex their right wrists whenever they wish and to report on when they first had certain conscious experiences – variously described as experiences of an urge, intention, or decision – to do what they did. After they act, they make their reports. Each participant performs many flexes during the course of an experiment and makes many "consciousness" reports. When participants are regularly reminded not to plan their wrist flexes and when they do not afterward say that they did some such planning, an average ramping up of EEG activity (550 milliseconds before muscle motion begins) precedes the average reported time of the conscious experience (200 milliseconds before muscle motion begins) by about a third of a second (Libet, 1985). Libet claims that decisions about when to flex were made – unconsciously – at the earlier of these two times (1985, p. 536). (The initial ramping that I mentioned is the beginning of a readiness potential, which has been defined as "a progressive increase in brain activity prior to intentional actions, normally measured using EEG, and thought to arise from frontal brain areas that prepare actions" (Haggard, Mele, O'Connor, & Vohs, 2015, p. 325). I return to readiness potentials in Section 12.4.)

Around 200 milliseconds (200 ms, for short) before muscle motion began, some of these participants may (occasionally) have had experiences of what they might describe as deciding to flex. And others, around that time, may instead have had experiences of something they might describe as having an intention to flex or, alternatively, an urge to flex. Now, we often have proximal urges that we do not act on, and sometimes we intend *not* to act on particular proximal urges. So, people might be capable of distinguishing some experiences of proximal urges from some experiences of proximal intentions. For example, they might be capable of distinguishing an experience of an urge to order another pint of beer from the waiter who is asking them for drink orders from the experience of an intention to do that. But how good are people at distinguishing experiences of deciding to do something, as I conceive of it, from experiences of a nonactionally acquired intention to do something?

In section 12.1, to illustrate the point that not all intentions are formed in acts of deciding, I highlighted an action that is part of a routine – unlocking my office door when I arrive there in the morning. If I had been asked a few seconds in advance to report on when I first became aware of a proximal intention to unlock it, I could have complied, and I could have reasonably reported that although I was aware of a proximal intention to do this, I was not aware of any proximal decision to do it.[11] (I can claim, consistently with my view of practical deciding, that because I was not at all uncertain what to do about the door, I made no decision about the door.) Matters are different in Libet's experiment. Participants are, for a time, unsettled about when to flex next.[12] Given

that that is so, can they distinguish between the experience of proximally deciding to flex, deciding being understood as an action, and the experience of nonactionally acquiring a proximal intention to flex – or between the former experience and becoming aware of an intention that was acquired a short time ago? Do such things "feel" the same? Do they "feel" different? I return to these questions shortly.

## 12.4 Picking

Participants in Libet's study have no reason to prefer any moment to any nearby moment to begin flexing. They are in what Edna Ullmann-Margalit and Sidney Morgenbesser call a "picking situation" (1977, p. 761). Ullmann-Margalit and Morgenbesser describe a "simple picking situation" as a "selection situation" in which the alternatives, A and B, are such that: (1) the agent cannot select both, (2) the agent is indifferent between them, and (3) the agent prefers selecting either of A and B to selecting neither (1977, pp. 757–758). There are less simple picking situations, of course: for example, the number of alternatives with the relevant features can be increased. And there are cases in which, although A and B are very different in value, agents will discover their value only after they select one of them. Think of a game show in which a contestant is invited to select either door A or door B, is informed that there is a new car behind one and an old goat behind the other, and has no evidence about which of the two doors hides the desired prize.[13]

Picking is one thing and reporting what one picked is another. The latter is an overt action, but the former, as I use the term "picking," is not. The same goes for, say, picking a key to press on a keyboard and pressing it. The latter is an overt action and the former is not. Suppose you are asked to pick a key to press on your laptop, to remember which key you picked, and to press it after counting to ten. Your picking a key to press is a mental event (and, more specifically, a mental action).

Carla is invited to pick a card – any card – from a deck of cards being held face-down by a magician. Her instructions are to pick a card to pull from the deck, to count to ten, and then to reach out and take that card. Carla is confident that she will arbitrarily pick a card, but she does not just haul off and pick one. Instead, she entertains various options for a card to pick and soon picks one.

We ask Carla whether she actively picked a card or instead nonactionally acquired an intention to pull that card out of the deck later. She asks us to explain our question, and we do. Carla replies that she actively picked that card – that she performed the mental action of picking it. We then ask how she can tell the difference between nonactionally acquiring an intention to do something and actively picking. Carla reports that there is a phenomenological difference. Under normal conditions, she says, when she gets to the peanut display in her local supermarket, she

just grabs one of the many jars of her favorite kind and puts it in her cart. In her view, she intends to grab that jar, but she does not actively pick it. In such situations, Carla reports, she lacks the feeling of active picking that she has in some other situations – for example, in the case of the magician's request. She describes the feeling as that of actively settling a question about what to do.

When we ask Carla whether she was uncertain or unsettled about what to do in the example she offered us, she says no. So her example of a nonactionally acquired intention resembles, in a certain respect, an example of mine that I commented on earlier: unlocking my office door. In neither case is there uncertainty or unsettledness about what to do. We ask Carla whether she would have had the feeling of actively picking a card if, as it happened, her unsettledness about which card to pick had been resolved by a nonactionally acquired intention to pick a certain card. Does resolution of this matter by a decision feel any different from resolution of it by a nonactionally acquired intention, we ask. Carla is silent; she is thinking. I return to Libet while she is mulling things over.

A brief description of my own experience as a participant in a Libet-style experiment (see Mele, 2009, pp. 34–36) may prove interesting. I had just three things to do: watch a fast clock with a view to keeping track of when I first became aware or conscious of something in the ballpark of a proximal urge, decision, or intention to flex my right wrist; flex whenever I felt like it (many times over the course of the experiment); and report, shortly after each flex, where I believed the hand was on the clock at the moment of first awareness. (I reported this belief by moving a cursor to a point on the clock. The clock was very fast; it made a complete revolution in about 2.5 seconds.) Because, for some time at the beginning of the experiment, I did not experience any proximal urges, decisions, or intentions to flex, I hit on the strategy of saying "now!" silently to myself just before beginning to flex. This is the mental event that I tried to keep track of with the assistance of the clock. I thought of the "now!" as shorthand for the imperative "flex now!" – something that may be understood as an expression of a proximal decision or intention to flex.

I definitely silently said "now!" to myself many times during the experiment. Those silent speech acts were mental actions. But did I make a proximal decision to flex (construed as a mental action) on any of those occasions? Did I only nonactionally acquire proximal intentions to flex? If the latter, were my silent speech acts prompted by the onset of those intentions? Or were they related to those intentions in some other way instead?

These questions are difficult to answer. One way to interpret my silent speech acts is as internally generated go signals. In go-signal studies, subjects know what they are supposed to do when they detect a go signal – for example, flex a wrist as soon as they can. Go signals in those studies

are externally generated: a beep is a good signal. My silent speech acts may be interpreted as internally generated analogs of such a beep.

Imagine a go-signal study in which subjects are instructed to flex their right wrists as soon as they can in response to a beep. According to one way of thinking about what happens, their detection of the beep prompts a proximal intention to flex, and the acquisition of that intention (or the neural realizer of that event) issues in a flex.[14] In a study of this kind, if the subjects behave as requested, there is no place for a decision about what to do when the beep is detected; for there is no uncertainty about what to do then.[15] If my silent speech acts function just as go signals are hypothesized to do in a study of this kind, they prompt proximal intentions to flex in the absence of any proximal decisions to flex. This leaves it open that I sometimes proximally decided to say "now!" But it also leaves it open that I never did and that my silent speech acts were prompted by nonactionally acquired proximal intentions to say "now!"

In Mele (2009), I suggested that some of the subjects in studies of the kind I participated in may "treat the conscious urge [to flex] as what may be called a *decide signal* – a signal calling for them consciously to decide right then whether to flex right away or to wait a while" (p. 75). Judy Trevena and Jeff Miller conducted a pair of interesting studies involving a decide signal (2010). Both studies had an "always-move" and a "sometimes-move" condition (Trevena & Miller, 2010, p. 449). In one study, participants in both conditions were presented with either an "L" (indicating a left-handed movement) or an "R" (indicating a right-handed movement) and responded to tones emitted at random intervals. In the sometimes-move condition, participants were given the following instructions: "At the start of each trial you will see an L or an R, indicating the hand to be used on that trial. However, you should only make a key press about half the time. Please try not to decide in advance what you will do, but when you hear the tone either tap the key with the required hand as quickly as possible, or make no movement at all" (Trevena & Miller, 2010, p. 449). The tone may be viewed as a decide signal calling for a proximal decision about whether to tap or not. (In the always-move condition, participants were always to tap the assigned key as quickly as possible after the tone.)

In a second study, Trevena and Miller left it up to participants which hand to move when they heard the decide signal. As in the first study, there was an always-move condition and a sometimes-move condition. In the always-move condition, the tone may be regarded as a decide signal calling for a proximal decision about which hand to move; and in the sometimes-move condition, it may be interpreted as calling for a proximal decision about whether to move at all and, if so, which hand.

Do participants ever respond to decide signals in these studies with decisions, as I conceive of them – momentary actions of intention formation? Do they respond instead only with nonactionally acquired

intentions? Participants in the always-move condition of the second study were given the following instructions: "When you hear the tone, please quickly tap with whichever hand you feel like moving. Please try not to decide in advance which hand you will use, just *wait for the tone and then decide*" (Trevena & Miller, 2010, p. 452; my italics). Notice that they are explicitly asked to *decide*. If participants were to understand practical deciding as depicted in view 1, they would take these instructions to call for them to nonactionally acquire an intention about which key to tap after they hear the tone and not before. And how would they do that? Their predicament would resemble mine as a subject in a Libet-style experiment. I took my instructions to call for me to wait for an urge to pop up – that is, more technically, to wait for a nonactionally acquired urge – and that was not happening. On the assumption that participants in the study at issue now understand their instructions in accordance with view 1, they would seem to take those instructions to call for them to wait for a relevant intention to pop up after they hear the tone; and they might find that this is not happening. If, alternatively, they understand their instructions in accordance with view 2, they regard their task as that of arbitrarily performing a momentary mental action of forming an intention about which key to press – that is, actively picking a key to press. Such an understanding does not encourage waiting for something to happen. Instead, it calls for a mental analog of a simple physical basic action – for example, raising a finger.

Operating with the actional interpretation of the instructions would make the participants' task more manageable. But it does not follow from this that practical decisions are actions. It may be that regarding decisions as actions is useful – or useful in certain circumstances – even if they are not in fact actions. And it may be that thinking of decisions as actions puts the participants in a position to acquire proximal key-tapping intentions quickly and nonactionally. Their efforts to decide which key to tap, as they conceive of deciding, may quickly issue in nonactionally acquired proximal intentions.

What would I try to do if I were a subject in Trevena and Miller's always-move decide-signal experiment? I would try to wait for the decide signal and then arbitrarily decide, as quickly as possible, on a key to tap. I would conceive of my arbitrary decisions as momentary actions of intention formation. How would I proceed if I were to accept view 1? I believe I would pretend that view 2 is true and conduct myself accordingly.[16]

## 12.5 More on views 1 and 2

Even Brian O'Shaughnessy, a proponent of view 1, reports that "we incline to the view that deciding-to-*do* is an activity" (1980, vol. 2, p. 297). Why do we have this inclination? Perhaps because we see the

view as expressing a plausible interpretation of our experience of practical deciding. But suppose that this plausible interpretation is incorrect and that view 1 is true. What philosophically interesting implications might that have?

As Randolph Clarke observes, "deciding or choosing … receives more attention than any other type of action in discussions of free will, particularly in discussions of libertarianism," the thesis that free will exists and is incompatible with determinism (2003, p. 23). He reports that "Libertarians have typically held that only mental actions – or mental actions of certain subclasses, such as decisions or choices, volitions, efforts, tryings, or willings – can be directly free, with the freedom of any actions of other types stemming from that of these mental actions" (Clarke, 2003, p. 121). If practical decisions are not actions, they must be eliminated from this list of alleged heavy lifters. (Clarke identifies decisions with choices; 2003, p. 125.)

This elimination from the list is a philosophically significant consequence of the truth of view 1, especially in light of Clarke's observation that practical decision is the item on the list that receives the most attention in the free will literature. This consequence provides some motivation to ask how strong the case is for view 1.

O'Shaughnessy seeks insight into the nature of practical deciding by examining cognitive deciding (1980, vol. 2, pp. 297–302). He argues that no cognitive decision is an action. But, of course, proponents of view 2 who accept this assertion about cognitive deciding will reject the claim that cognitive deciding provides a good model for practical deciding.

O'Shaughnessy contends that a nonactional view of practical decisions that parallels his view of cognitive decisions "receives confirmation in the fact that there is no order: 'Decide to raise your arm'" (1980, vol. 2, p. 300). This is unconvincing. Return to Joe and Jill. When, after telling Jill that he has decided yet again that it would be best to quit smoking, Joe reports that his next hurdle is to decide to quit, we can easily image Jill saying, "Well, then, decide to quit!" She might add the following, if she has a philosophical streak: "If I had the authority to command you to do things, I would command you right now to decide to quit."

O'Shaughnessy presents a case in which "a jury of one is trying to decide whether or not to bring in a verdict, Guilty" (1980, vol. 2, p. 300). This juror's sole concern is whether "it is certain that the defendant committed the crime" (italics eliminated). O'Shaughnessy contends that the juror's deciding that the person is guilty is "distinct from" his deciding to issue a verdict of guilty and that both decisions are nonactional (p. 301). The former decision, he contends, is the nonactional acquisition of a belief that the defendant is guilty while the latter is the nonactional acquisition of an intention to issue a verdict of guilty.

A proponent of view 2 sees things differently. In this case, the juror's coming to the conclusion that the defendant is guilty settles for him the question what to do. What we have here, according to a proponent of view 2, is a case in which the cognitive decision issues immediately and by default in an intention to pronounce the person guilty; the juror makes no decision to make this pronouncement. Once the juror is convinced that the person is guilty, he is no longer uncertain about what to do. (Bear in mind the juror's sole concern.) At this point, according to a proponent of view 2, there is no place for a decision to issue a guilty verdict. O'Shaughnessy does not provide compelling grounds for rejecting this view of his story about the lone juror.

O'Shaughnessy contrasts his view of "the relation between" cognitive and practical deciding with "the account of the relation propounded by believers in the Gide-ean *acte gratuit*" (1980, vol. 2, p. 300). Such believers are represented as maintaining that "a deed of 'pure freedom'" is done for "*no* reason" and must "emerge out of the blue." These believers, O'Shaughnessy observes, may easily detach practical decisions from cognitive ones. However, the case I presented of an agent with a practical question that is not settled by his cognitive decision about the matter is much more mundane. Even though Joe decides that it would be best to quit smoking, beginning tonight, he continues to be unsettled about whether to do that. If he decides to keep smoking, he will do that for reasons. And the same is true if he decides to quit. There is no Gide-ean *acte gratuit* here.

Do neuroscience experiments provide good grounds for believing that there are no practical decisions, as I conceive of such decisions? Consider the following claim: (*CN*) Libet-style experiments show (1) that we never make decisions, as construed on view 2, and (2) that instead we nonactionally and unconsciously acquire pertinent intentions. Although I have had a lot to say about the experiments at issue (references include Mele, 2009, 2018), I have never taken up precisely this claim. If actional decisions are at the core of free will, *CN* may certainly be presented as a challenge to believers in free will.

Fortunately for me, objections I have raised to assertions similar to *CN* apply also to *CN* itself. As I mentioned, I have argued that we do not have good evidence for the claim that pertinent decisions are made "early" in the experiments at issue and that we have better evidence that if pertinent decisions are made, they are made around the time the participants say they make them (Mele, 2009, 2018). My arguments for this also apply to parallel claims that substitute early nonactionally acquired intentions for early decisions; the arguments, therefore, apply to *CN*. A Libet-style case for the truth of *CN* would rest on the claim that pertinent intentions arise at the early times at issue, before participants are conscious of them.

Another point I have made elsewhere is that we cannot safely generalize from alleged findings in picking situations to certain claims about

all practical decisions (Mele, 2009, pp. 83–87; 2013, pp. 4–7; 2018, pp. 380–381). In Libet-style studies (Libet, 1985; Soon et al, 2008; Fried, Mukamel, & Kreiman, 2011), participants arbitrarily pick something – a moment to flex a wrist or tap a key, which of two buttons to press, and so on. Even if it were shown that decisions, as I conceive of them, are not made in such picking situations and that, instead, intentions are nonactionally and unconsciously acquired, we would not be justified in concluding that this is what happens in all situations that may be claimed to involve actional practical deciding. If it were shown that nonactionally, unconsciously acquired intentions break ties when an agent's attitude toward pertinent options is one of indifference, it would not follow that this is what happens in situations like Joe's, situations that feature conscious reasoning about what to do and conscious conflict regarding options with respect to which the agent is far from indifferent. Perhaps such features help pave the way for deciding that is both conscious and actional. Furthermore, we now have some evidence that different neural processes underlie arbitrary picking, on the one hand, and what the experimenters call "deliberate decisions," on the other – for example, decisions (or apparent decisions) made about which of two charities to donate to when one likes them both, is not indifferent between them, and is unsettled about which one to donate to this time (Maoz, Yaffe, Koch, & Mudrik, 2017). This evidence constitutes an additional problem for the generalization at issue.

I should make it clear that I am not conceding that decisions, construed as actions, are not made in picking situations. In my opinion, it is an open question whether they are. Attention to another experiment sheds light on this openness.

Aaron Schurger, Jacobo Sitt, and Stanislas Dehaene contend that the brain uses "ongoing spontaneous fluctuations in neural activity" (2012, p. E2904) – neural noise, in short – in solving the problem about when to act in Libet-style studies. A threshold for a "neural decision" is set (p. E2904), and when such activity crosses it, a neural decision is made. They contend that most of the readiness potential (described in Section 12.2) – all but the last 150 to 200 ms or so (p. E2910) – precedes that decision. In addition to reporting evidence for this that comes from the work of other scientists, Schurger and colleagues offer evidence of their own. They use "a leaky stochastic accumulator to model the neural decision" made about when to move in a Libet-style experiment, and they report that their model "accounts for the behavioral and [EEG] data recorded from human subjects performing the task" (p. E2904). The model also makes a prediction that they confirmed: namely, that when participants are interrupted with a command to move immediately (press a button at once), short response times will be observed primarily in "trials in which the spontaneous fluctuations happened to be already close to the threshold" when the command (a click) was given (p. E2905).

What led participants to pick the moment they picked for a button press, if such picking happened? The answer offered by Schurger and colleagues is that random noise crossed a neural decision threshold then. And they locate the time of the crossing very close to the onset of muscle activity – about 100 ms before it (pp. E2909, E2912). They write: "The reason we do not experience the urge to move as having happened earlier than about 200 ms before movement onset [referring to Libet's finding about consciousness reports] is simply because, at that time, the neural decision to move (crossing the decision threshold) has not yet been made" (Schurger, Sitt, & Dehaene, 2012, p. E2910).[17]

What exactly is a proximal *neural decision* to press a button in studies of the kind at issue and how is it related to personal-level events? Is the neural decision a cause of a conscious proximal decision to press, the decision being an action? Is it instead a cause of a nonactionally acquired conscious proximal intention to press? Is it a neural realizer of the former – or a neural realizer of the latter? Is it a cause – or neural realizer – of a personal-level decision (understood as an action) or of a nonactionally acquired intention that the person would – or might – not have been conscious of if he had not been prepared to report on a pertinent conscious experience? These questions are difficult. I know of no good reason to believe that view 1 comes out looking better than view 2 in this sphere.

My primary reason for discussing various neuroscientific experiments in this article was to help motivate a worry about my own stance on practical decisions. In arguing for the existence of practical decisions as I conceive of them, I appeal (in Mele, 2000 and elsewhere) to ordinary experiences of practical decision-making. But if the experience of actively making a decision in response to uncertainty about what to do is no different intrinsically from the experience people would have at the time if they were nonactionally acquiring a pertinent intention in response to such uncertainty, our experiences of decision-making are not much help in a dispute between proponents of view 1 and proponents of view 2 about the nature of our actual practical decisions.[18] Even so, the extant arguments for view 1 are weak, and an actional conception of practical deciding is much more consonant with ordinary thought and talk about practical deciding. So if I had to choose between views 1 and 2, I would continue to opt for view 2.[19]

## Notes

1. I do not assume that all mental actions are conscious actions. On this, see Mele (2010).
2. Brian O'Shaughnessy defends a view of this kind (1980, vol. 2, pp. 300–301). Also see Williams (1993, p. 36).

3. John McGuire (2016) has argued that in scenarios involving side-effect actions of a certain kind, one may decide to perform a side-effect action but not intend to perform it. For my purposes in this article, it can be left open that an exception should be made in the case of some side-effect actions.
4. For a strange case that seemingly provides an exception to this, see Mele (2000, pp. 90–91).
5. On akratic failures to intend, see Audi (1979, p. 191), Davidson (1980, ch. 2; 1985, pp. 205–206), Mele (1992, pp. 228–234; 2012, pp. 25–28), and Rorty (1980).
6. Proximal intentions also include intentions to continue doing something that one is doing and intentions to start *A*-ing (e.g., start running a mile) straightaway.
7. This is not to say, for example, that I was certain that I would unlock my door, if such certainty entails having an explicit belief that I will unlock my door. Compare "He was not uncertain about X" with "Y was not unpleasant for her." The latter sentence does not entail that Y was pleasant for the person; Y might have been neither pleasant nor unpleasant for her. Similarly, a person may be neither certain nor uncertain about X. I am in this condition regarding propositions I have never entertained, for example. In any case, as I approach my office door under normal conditions, I seem not to be consciously entertaining the proposition that I will open it.
8. Someone who uses "decide" in such a way that Henry counts as having decided to order the Cobb salad may or may not believe that all practical decisions are actions.
9. For discussion and references, see Mele (2009, 2018).
10. Readers should not infer that I place a lot of weight on these reports. There are grounds for doubt about the accuracy of the reported awareness times in these studies. I have discussed such grounds elsewhere (Mele, 2009, ch. 6; also see Maoz et al, 2015, pp. 190–194).
11. To say that I could have complied in this situation is not to say that I am aware of my intention to unlock my door in normal circumstances, which include my not being asked to report any intentions.
12. Being unsettled about when to do something differs from not being settled about when to do it. Let X be a course of action that Joe has never thought about engaging in – say, flying to Brussels. Is Joe settled about when to fly to Brussels? Obviously not. But he is not unsettled about this either. And, of course, rocks are neither settled nor unsettled about anything.
13. For comparable examples, see Ullmann-Margalit & Morgenbesser (1977, p. 764).
14. An alternative possibility is that the combination of subjects' conditional intentions to flex when they detect the beep and their detection of the beep initiates a flex in the absence of any proximal intention to flex. If and when that happens, it is false that a proximal intention to flex is produced by the brain before the mind is aware of it. This, of course, is contrary to Libet's interpretation of his results. On this, see Mele (2009, p. 62).
15. Here, obviously, I am assuming that deciding to *A* depends on uncertainty about what to do.
16. The same goes for Trevena and Miller's other decide-signal experiments. For discussion of their findings, see Mele (2018, pp. 376–377).
17. I discuss Schurger et al (2012) at greater length in Mele (2018, pp. 377–379). I have drawn on that discussion here.
18. I could have raised this worry without even mentioning neuroscience, of course. But the neuroscientific background I have discussed makes the worry more salient and, I believe, more interesting.

19. I led a seminar on a draft of this chapter at La Universidad de los Andes (September, 2017). I am grateful to the audience, especially Santiago Amaya and Sam Murray, for discussion. I am also grateful to an audience at the Institute of Philosophy, University of London (November, 2017) for discussion and to Sarah Paul for written comments on the draft on which my presentation was based. I am grateful as well to an audience at the University of Helsinki (August, 2019) for feedback and to Michael Brent for comments on the penultimate draft. This chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed here are my own and do not necessarily reflect the views of the John Templeton Foundation.

## References

Audi, R. (1979). Weakness of will and practical judgment. *Noûs*, *13*, 173–196.

Audi, R. (1993). *Action, intention, and reason*. Ithaca, NY: Cornell University Press.

Clarke, R. (2003). *Libertarian accounts of free will*. Oxford: Oxford University Press.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1985). Replies to essays I–IX. In B. Vermazen, & M. Hintikka (Eds.), *Essays on davidson*. Oxford: Clarendon Press.

Fried, I., Mukamel, R., & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*, *69*, 548–562.

Haggard, P., Mele, A., O'Connor, T., & Vohs, K. (2015). Free will lexicon. In A. Mele (Ed.), *Surrounding free will* (pp. 319–326). New York: Oxford University Press.

Kaufman, A. (1966). Practical decision. *Mind*, *75*, 25–44.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*, 529–566.

Libet, B. (2004). *Mind time*. Cambridge, MA: Harvard University Press.

Maoz, U., Mudrik, L., Rivlin, R., Ross, I., Mamelak, A., & Yaffe, G. (2015). On reporting the onset of the intention to move. In A. Mele (Ed.), *Surrounding free will: Philosophy, psychology, neuroscience* (pp. 184–202). New York: Oxford University Press.

Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2017). Neural precursors of decisions that matter – An ERP study of deliberate And arbitrary choice. *bioRxiv*. doi:. Retrieved from https://doi.org/10.1101/097626.

McGuire, J. (2016). Can one decide to do something without forming an intention to do it? *Analysis*, *76*, 269–278.

Mele, A. (1987). *Irrationality*. New York: Oxford University Press.

Mele, A. (1992). *Springs Of action*. New York: Oxford University Press.

Mele, A. (2000). Deciding to act. *Philosophical Studies*, *100*, 81–108.

Mele, A. (2003). *Motivation and agency*. New York: Oxford University Press.

Mele, A. (2009). *Effective intentions*. New York: Oxford University Press.

Mele, A. (2010). Conscious deciding and the science of free will. In R. Baumeister, A. Mele, & K. Vohs (Eds.), *Free will and consciousness: How might they work?* (pp. 43–65). New York: Oxford University Press.

Mele, A. (2012). *Backsliding: Understanding weakness of will*. New York: Oxford University Press.
Mele, A. (2013). Free will and neuroscience. *Philosophic Exchange*, *43*, 1–17.
Mele, A. (2017). *Aspects of agency*. New York: Oxford University Press.
Mele, A. (2018). Free will and consciousness. In D. Jacquette (Ed.), *Bloomsbury companion to the philosophy of consciousness* (pp. 371–388). London: Bloomsbury.
O'Shaughnessy, B. (1980). *The will*, vol. 2. Cambridge: Cambridge University Press.
Rorty, A. (1980). Where does the akratic break take place? *Australasian Journal of Philosophy*, *58*, 333–346.
Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, *109*(42), E2904–E2913.
Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545.
Trevena, J., & Miller, J. (2010). Brain preparation before a voluntary action: Evidence against unconscious movement initiation. *Consciousness and Cognition*, *19*, 447–456.
Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and choosing. *Social Research*, *44*, 757–785.
Williams, B. (1993). *Shame and necessity*. Berkeley, CA: University of California Press.

# 13 Self-control, Attention, and How to Live without Special Motivational Powers

*Sebastian Watzl*

## 13.1 Introduction

The image of the agent who controls her passions like a charioteer steering her hot-blooded horses, has long sparked the philosophical imagination.[1] Such self-control illustrates, it has been argued, something deep and interesting about the mind. Specifically, it has been suggested that for its explanation we must posit special motivational powers: *willpower* as an irreducible mental faculty (Holton, 2009), the active self as a dedicated and depletable pool of *psychic energy,* or – in today's more respectable terminology – mental resources (Baumeister, Bratslavsky, & Muraven, 2018), or a deep division between *reason and passion* – a deliberative and an emotional motivational system (Sripada, 2014).

This essay argues that no such special motivational powers are necessary. Yet, at the same time, the tradition is right that self-control powerfully illustrates the importance of a feature of the mind. What it illustrates, I argue, is the importance of the mental activity of attention in the control of *all* action. It is by appeal to this mental activity that we can dispense with special motivational powers.

The significance of attention for self-control, of course, is compatible with several models, including ones I would like to reject.

On the one hand, one might link attention to a willpower or a resource-driven view of self-control, starting perhaps from William James (1890, p. 562) who thought that "effort of attention is … the essential phenomenon of will". The idea might be that attention either is itself a mental resource, a willpower faculty, or the mechanism controlling access to them.

On the other hand, and in stark opposition to the first model, attention has also been linked to the denial that the will is involved in self-control at all, as what dissolves the charioteer altogether (Ganeri, 2017). Recently, attention has been suggested as central to a surprisingly non-agential view where all strategies for self-control are "distinctively cognitive" (Kennett, 2001, p. 139; cf. Kennett & Smith, 1996, 1997).

DOI: 10.4324/9780429022579-14

The role of attention I will argue for is opposed to both of these conceptions. The capacity for attention is important for self-control exactly because it is agential, and it is important for self-control even though it has no connection to anything resembling special motivational powers. The interesting feature of the mind self-control illustrates is this: attention acts as a flexible interface between the agent's motivational systems and her actions. Through attention an agent can actively couple or decouple an intention, preference, or desire to and from action – by intentionally changing the current priority of her mental states. I call the resulting view *the re-prioritization account* of synchronic self-control.

In one sense, the project of this paper is deflationary: self-control uses no special motivational capacities. Agents use attention for action control also when things go smoothly and when there is no need for controlling any wayward temptation. Once we think correctly about the role of attention in the control of all forms of agency (see Watzl, 2017; Wu, 2016), there is no explanatory role for willpower, mental resources, or a divided mind. Self-control is not special: if we think of Humeanism as the view that there is fundamentally only one kind of motivational system and that all action is based in that system, then this essay contributes to a defense of Humeanism.

In another sense, the project – though deflationary – is constructive. Any model of agency in terms of only beliefs and desires, motivational and representational states, or preferences and credences, is incomplete. Models of agency need attention as an independent factor. The way attention organizes the mind cannot be subsumed under what the agent wants and how she takes the world to be.[2] A different conception of Humeanism as the view that every mental state is either motivational, representational, or a combination of them, is false.

The view I argue for in this essay aligns with one defended by Inzlicht, Berkman, and colleagues in the psychological literature: according to them "the decisions that we label self-control are merely a fuzzy subset of all value-based decisions, which involve selecting a course of action among several alternatives." (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017b, p. 423). I also agree with them that the field "would be better served by abandoning the resource concept altogether" (Inzlicht & Berkman, 2015, p. 520), and that "self-control outcomes emerge organically from the operation of a single, integrative system with input from multiple regions rather than antagonistic competition between two processes" (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017b, p. 424). I point readers with an interest in the empirical details to this literature. My goal here is in a corresponding philosophical defence.

My primary aim is descriptive: to show how self-control can work without willpower, mental resources, or mental division. In the conclusion,

I will briefly touch also on normative questions: whether self-control can be rational or irrational, and whether exercising self-control is always good for the agent. If my descriptive model is right, then, there probably isn't much, in general, to say about the normativity of self-control. The descriptively deflationary view may help to show that also in the normative sense we should neither take an overly elevated nor a debased view of self-control.[3]

Here is how I will proceed. In Section 13.2, I will introduce what self-control is, and focus on one variant that philosophers have found especially puzzling. In Section 13.3, I explain one such philosophical puzzle, aiming to show that this variant, intentional synchronic self-control, is impossible without positing special motivational powers. A satisfactory theory of self-control should respond to this puzzle. In Section 13.4, I will first sketch my re-prioritization view of self-control and show why a non-intentional account of synchronic self-control, which shares some of its features, is unsatisfactory. In Section 13.5, I present the view in a bit more detail and show how it can be used to defend the possibility of intentional synchronic self-control on the basis of four premises. Sections 13.6, 13.7, 13.8, and 13.9 then defend each of these premises. Section 13.10 answers how the re-prioritization view can explain the sense of effort and difficulty accompanying self-control attempts, and why self-control may improve through training. In the concluding Section 13.11, I summarize and point briefly to potential normative implications.

## 13.2 What is self-control?

Agents exercise self-control to counteract a threat of losing control over themselves. Here is a paradigmatic example:

> *Cookie Temptation.* After an exhausting day at work, on her walk home, Christina walks by a pastry store. Delicious cookies are on display. Christina stops and looks into the window. The temptation rises: she wants one. Eat it. Right here. Right now. But Christina has resolved to become more fit and healthy, and made the specific plan to go on a run first and then eat a salad. She firmly believes that this is what she should do. Part of her resolution was specifically to overcome temptations like the one she is now facing.

When Christina is standing in front of the pastry shop, she experiences a threat of losing control over herself. She feels that her momentary urge to eat that cookie is getting the better of her. But this is not what she thinks she should do. In a situation like this, she may exercise self-control: she may pull herself together and go on the run she had planned.

Here is another paradigmatic case:

> *Angry Punch.* Amira is driving home from a dinner with friends. Suddenly, a police car comes up behind him, and pulls him over. When police ask him for his license, he hears in their voice that they treat him as some kind of suspect. Then they interrogate him about whether he had taken drugs, ask him to step out of the car, and begin searching every corner inside. Amira feels the anger rising. He feels insulted by their demeanor. He feels like punching one of the policemen in the face. But he knows that this would have disastrous consequences. He would probably end up in prison or worse.[4,5]

For Amira, like for Christina, there is a threat of losing control over himself. He believes that no good is going to come of that punch. He needs self-control to do as he thinks he should do: let the insulting procedure pass, get back into the car once it is over, and be home in less than ten minutes.

Starting from such paradigmatic examples, we can characterize self-control in terms of two features (cf. Duckworth, Gendler, & Gross, 2016):

The first feature is a certain type of situation, a *loss of control threat*, as I will call it. Loss of control threats are characterized by an asymmetric and subjective conflict. The agent is faced with at least two options. One of the options is, from her own perspective, *better*. The other one, though, has a more *powerful grip* on her. In the cookie temptation case, the better option is to go on a run and eat a salad. The option with the more powerful grip is to eat a cookie right there on the spot. The conflict in the situation is subjective in three ways. First, the agent is aware of both options. Second, it is from the agent's own perspective that the two options are incompatible. Third, the better option is better by the agent's own lights. Loss of control threats should be distinguished from situations where the agent is not aware of an important option, ones where the agent does not realize that there is a conflict between what has a powerful grip on her and what she takes as the better option (a form of mental fragmentation), cases where the agent is stuck between two equally desired options (Buridan's ass situations), and cases where one option is better for the agent even though the agent herself does not see it that way.

The second feature of self-control is conflict resolution: the agent counteracts the threatening loss of control or – in other words – she resolves the subjective conflict in favor of the better option. Christina would not have exercised self-control if she went home for her run because the cookie shop is closed, or because her cruel partner does not allow her to eat the cookie. For a case of self-control, the resolution of the conflict must be causally traced to the agent. This, though, is not sufficient: Christina would also not have exercised self-control if she did not eat the cookie because she suddenly feels a stomach ache, or if she

happens to remember her overdue taxes whose urgency crowds out even the cookie. Self-control must be agential, in that the conflict is resolved by something the agent herself does non-accidentally.[6]

I will speak of a *self-control event* when an agent, in the non-accidental sense just described, averts a loss of control threat. A *self-control strategy* is a way of averting loss of control threats. And an agent has the *capacity for self-control* insofar and to the degree to which she is competent at averting such loss of control threats.[7]

Given this characterization of self-control, we can distinguish types of self-control by varying self-control cases along various dimensions.

First, there are different ways of understanding when an option is *better* from the agent's own perspective: it is the one that is more congruent with the agent's more long-term goals; more congruent with the agent's overall motivational set; the one that the agent more closely identifies with; the one that the agent would choose if she were fully rational; the one that she has formed a resolution or intention to pursue; or, finally, we may think of an option as subjectively better because it is the one the agent believes or judges to be the option she should pursue. I will here focus on options that are subjectively better in this last sense.

Second, there are also different ways of understanding when an option has a *more powerful grip* on the agent: it may, for example, be the one that promises the more immediate reward, or the one that is phenomenally more powerful. My discussion will focus on options whose powerful grip consists in the fact that the agent in the subjective conflict situation prefers the option over the one she judges to be the one she should pursue (we may also say that she desires the option more).

Third, there are different types of self-control strategies. On the one hand, we have diachronic strategies, which resolve the subjective conflict before it arises. The agent may, for example, eat a banana before leaving the office knowing that this will help her not succumb to eating cookies on her way home. On the other hand, we have synchronic strategies, which resolve the conflict once it has arisen. A second distinction is between situational strategies, which are pursued by selecting or modifying situations so as to minimize or resolve potential conflict – e.g., by walking home on a route that avoids the pastry store (cf. Duckworth et al., 2016), and intra-psychic strategies, which are pursued by changing one's own mind. I will focus on synchronic and intra-psychic strategies.

In what follows, I will thus understand self-control events as cases where an agent non-accidentally changes her own mind in such a way as to bring her actions in line with her judgment about what she should do when faced – at that time – with an opposing preference. The relevant self-control strategies are ways of changing her mind in such a way. The relevant capacity for self-control is an agent's competence in achieving such a change of mind. The threat of losing control here thus consists in a threat of akrasia, i.e., acting against one's best judgment. By engaging

in self-control, the agent faces that threat and does what she thinks she should do. In a slogan: self-control here is the enkratic aversion of a threat of akrasia.

It's easy to see why this type of self-control might suggest that agents have special motivational powers. If the horses are the agent's preferences or desires, which pull her in various directions, then the enkratic aversion of akrasia seems to show that the agent herself – *as opposed to* her preferences – can get the horses to do what *she thinks is right*. But then she must have some special power – her self, her will, her reason – to align her preferences with her beliefs. There must be more than horses.

My own view of self-control is that, as far as motivation is concerned, there is no more than horses. No variety of self-control is resolvable only through special mental capacities. I choose to focus on the enkratic aversion of akrasia because this is the version that seems to raise a philosophical puzzle.

## 13.3 The preference determines action principle and the paradox of self-control

The puzzle is that this type of self-control, while intuitively quite common, can appear to be impossible given a relatively plausible assumption of how the mind works. This so-called *paradox of self-control* (cf. Kennett & Smith, 1997; Mele, 1992; Sripada, 2014) has inspired special motivational powers as well as the non-agential view of self-control mentioned in the introduction.

Consider the following plausible principle:

> *Preference Determines Action*. If, at time *t*, the agent prefers option *A* over option *B*, and believes at *t* that both *A* and *B* are genuine options for her, then the agent will take option *A* (and not *B*) – given that she chooses one of them intentionally.

The paradox can now be presented in the form of the following dilemma (Sripada, 2014). I will illustrate it with the Cookie Temptation case:

> Horn 1: Suppose that when in front of the pastry store, Christina indeed prefers eating the cookie to going home and preparing for her run. It follows from the fact that we have a subjective conflict situation that she recognizes eating the cookie as one of her genuine options. But then it follows from the Preference Determines Action principle that if she chooses either eating the cookie or preparing for her run intentionally, she will choose eating the cookie. But this then seems to rule out that Christina will engage in another activity, self-control, that will make her *not* eat it.

> Horn 2: But then suppose, by contrast, that Christina's tempting urge does not reflect a genuine preference at that time for the cookie over the run. Or suppose that Christina doesn't actually take herself to be free to eat the cookie, i.e., that she doesn't recognize it as a genuine option for her at that time. Then, self-control would seem to be superfluous: if she doesn't actually want the cookie (or prefer it over the run), or thinks that she cannot actually take that option, then she doesn't need to exercise self-control to prevent herself from eating it.

The paradox is that the exercise of synchronic self-control can seem either impossible or superfluous. A philosophically satisfactory account of self-control needs to show how enkratic aversions of threats of akrasia are possible and efficient at the same time. If, as we have stipulated, the agent indeed prefers a course of action that she believes is open to her, then how could she intentionally do something else that is in direct conflict with what she prefers? In the chariot image: if preference determines action, then the only thing that can change the direction of the chariot are the horses. But then if the horses pull left, it just isn't possible that the chariot ends up going right – unless, of course, it happens accidentally (a gush of wind blows against the horses) or there is a charioteer with special steering powers.

I said that the Preference Determines Action principle is plausible. What exactly is its status?

On one construal, the principle is analytic, following from the definition of "preference". According to some revealed preference theories, choice behavior determines preferences: from the fact that an agent takes option B over option A, in a situation where both options are subjectively available to her, we can infer that she prefers B over A (cf. Samuelson, 1938). If this were right, then we would know that Christina, when she "succeeds in self-control" and pursues Run (B) over Cookie (A), must have preferred B over A, and hence our original description in terms of a preference of A over B must have been wrong.[8]

Given appropriate consistency constraints, choice behavior can without doubt be used to define a preference ordering for the agent (as the so-called revelation theorems show). But the fact that we can *define* such an ordering doesn't show that this ordering is explanatory or psychologically real: preferences should be construed as mental states that *explain* behavior not as summaries of such behavior (cf., e.g., Hausman, 2000; Bermúdez, 2009; Dietrich & List, 2016).[9] The revealed preference paradigm, while influential, is arguably inspired by a form of behaviorism we have little reason to accept (cf., Hausman, 2000; Dietrich & List, 2016). Choice behavior can be *evidence* for preferences without determining them.

I therefore do not take the Preference Determines Action principle as following from the definition of preference. Rather, I take it as

a substantial and explanatorily powerful hypothesis about intentional action. It seems to be deeply embedded in our folk psychology, a "truism" (Kennett & Smith, 1996, 1997), and it accords well with modeling in, for example, the economic sciences. A theorist who rejects the principle will need to show why it fails where it fails and how it needs to be amended. Such rival explanatory schemes will need to be judged by whether they are better than the one promised by the principle itself.

I will show that there is such a better explanatory scheme.

## 13.4 Attention and the non-agential view

In my view, synchronic self-control strategies are ways of re-focusing attention, or re-prioritizing mental states, as I will call it. The capacity for self-control, accordingly, consists in a skilled competence at such re-prioritizing in situations where the agent is faced with a contrary preference. What the Preference Determines Action principle leaves out is that attention acts as a flexible interface between preferences and actions. An agent's distribution of attention mediates between the agent's motivational systems and her sensory situation. Because their influence is so mediated, the agent can intentionally decouple her preferences from the relevant action. This is what she does when she intentionally engages in synchronic self-control.

Before I present my own re-prioritization view of self-control, let me briefly look at one that looks similar but promises to keep the Preference Determines Action principle untouched. The view is Jeannette Kennett's (2001) (see also, Kennett & Smith, 1997). Here is how she puts it:

> When an agent realizes that her actual desires do not match her judgements of desirability, and that she is therefore in danger of losing control of what she does, there are three ways in which she may focus her attention so as to bring it about that she does as she believes she should. First, she may restore the focus of her attention. Second, she may narrow or redirect the focus of her attention. Third, she may expand the focus of her attention.
>
> (Kennett, 2001, p. 136)

Kennett here seems to agree that self-control is achieved through strategies for re-focusing attention.

But *how* do re-focusing strategies work? What is distinctive of Kennett's view is that she thinks that the relevant attention shifts are never intentional actions (they can't be, for Kennett, because of "truisms" like the Preference Determines Action principle). Re-focusing of attention, for Kennett, then, is always "distinctively cognitive" or "a matter of her entertaining or excluding certain thoughts at the appropriate time" (op. cit., p. 139). The relevant changes of attention thus are not explained

on the basis of the agent's motivational states. Therefore, there is no motivational conflict with the motivation associated with her preference.

One challenge for the non-intentional view is to explain how the relevant un-motivated changes are interestingly different from an accidental avoidance of the loss of control threat. You don't eat the cookie because of a sudden stomach ache, and a sudden thought about your looming taxes prevents you from punching a policeman. The fact that these new "thoughts" occupy your attention at just the right time will, as a matter of fact, avoid the loss of control threat. But these are not instances of self-control.

This specific challenge, arguably, can be answered. For example, one might say that the relevant attention shift, while not an intentional action, must be the result of a reliable competence, a subject-level dispositional capacity. In the stomach ache and urgent tax examples, the relevant attention shifts are not the result of such a reliable competence. Maybe we could even call the deployment of such capacities an agent's "activity" (cf. Schellenberg (2019) on perceptual consciousness as an activity).

This still leaves the feeling that something has been left out: the agent's choice of the better option seems intentional and motivated in a way that perceiving a red dot (one of Schellenberg's examples) is not: maybe the agent deploys reliable capacities also in the latter case, but perceiving still seems to be just happening to the agent: she has no voluntary control over it, and cannot perceive intentionally.

We can sharpen this problem: the restoring of the focus of attention, or the narrowing, re-directing, or expanding of it, all – unlike perception, and unlike having thoughts – are exercises of agency: attention shifts, as I argue in detail in Watzl (2017), are always based in the agent's motivational system. While it may be true that "people exercise control over their own thought processes simply by having the thoughts that they are disposed to have" (my emphasis), as Kennett and Smith (1997, p. 124) put it, people control their attention like they control their (other) actions: try listening to the subtle flute in the big orchestra and keep your attention on it. This is an action you can control like you can control the movements of your fingers. Given that attention can be shifted intentionally, an unintentional attention shift seems accidental. The description in terms of attention arguably makes Kennett's view seem unlike the case of the thought about the looming taxes. But it is illegitimate, since it smuggles in an agential element that the view officially disavows.

The re-prioritization account, I will now argue, can explain self-control through attention while fully acknowledging its agential character.

## 13.5 The re-prioritization account of self-control

What is important for the role of attention in self-control is not that the relevant shifts of attention are unmotivated. What is important is how attention affects the agent's mind. Attention organizes mental states

in an action-relevant way: how much of an agent's attention a mental state occupies makes a difference to how that state influences the agent's actions (I take it that this is also a truism of folk psychology). By intentionally changing how much of her attention a mental state occupies, the agent can therefore intentionally change her course of action. Based on this simple idea, I argue that an agent can through an intentional change to her distribution of attention intentionally prevent herself from acting on a preference, and thus intentionally engage in self-control. This self-control strategy, I argue, requires no special motivational powers, and explains everything about self-control that needs explaining.

The attentional account of self-control draws on both the shifting priorities model recently proposed in the empirical literature by Michael Inzlicht and others[10], as well on philosophical accounts of attention and its role for action provided by Wayne Wu (2014, 2016) and myself (Watzl, 2017). I will provide a philosophically satisfactory and elaborated account that shows how "[a]ttention plays a crucial role in … self-control by gating which options enter the choice set at any one moment and foregrounding their salient attributes" (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017b, p. 423).

I will first illustrate the basic idea. As a starting point ask how a preference actually brings about an action. The way preferences bring about intentional actions, I argue, is through the agent's distribution of attention: attractive, affectively loaded or action-relevant features of the relevant option will be highly salient to her, drawing her attention to them. Further, her preference directs her attention to relevant targets for action: from all possible targets for action, Christina must actually find the cookie before she could eat it. Intentional action, in most cases, requires an alignment of the distribution of attention with one's preferences.

But since preference influences action through the agent's distribution of attention, I argue further, the agent can intentionally interfere at this stage. Importantly, in doing so the agent does not act from any motivation that does not derive from her preferences. It is compatible with a preference of A over B that the agent prefers that the attractive, action-relevant properties of A and the objects targeted by A-ing occupy less of her attention than they actually do (in a self-control situation, the agent may have that latter preference alongside the former because she believes that B and not A is the better option).

But if the agent prefers that certain properties of A or objects targeted by A-ing occupy less of her attention than they currently do, then she can, based on that preference, intentionally shift her attention away from those properties or targets. This can break the causal link between the preference and the action (decoupling the action from the preferences) and so the agent will end up not acting on her preference of A. This in turn will often lead to a preference change so that she now starts

preferring B over A. And so, Christina ends up going on her run, and Amira back in his car – with no special motivational powers needed.

In the following sections, I present this idea in terms of a detailed argument based on four premises.

My presentation of this argument will use some aspects of the idea, defended in Watzl (2017), that attention is a form of mental agency where agents act on their own minds by changing or maintaining the priority ordering of their occurrent, subject-level mental states. According to this *priority structure view*, attention consists in the agent's activity of regulating priority structures, which order the parts of the subject's current state of mind by their current priority to the subject: when an agent is perceptually attending to a perceptually presented item or a perceptually presented feature, she is prioritizing some parts of her overall perceptual state over other parts of that state. If attention is, for example, visually focused on an object, then the state of seeing that object is prioritized over other parts of that visual state. When attention is focused on a feature like the color of an object, then the state of seeing that feature is prioritized. The priority structure view also allows for non-perceptual forms of attention. These are at issue when, for example, Kennett speaks of attention as being "a matter of … entertaining or excluding certain thoughts" (op. cit., p. 139). Attention can bring occurrent thoughts to the subject's mind and prioritize them over other parts of her current mental state. When attention is non-perceptual what is prioritized thus is a non-perceptual aspect of the subject's on-going mental life: this may, for example, be a thought, a bodily sensation or a feeling, or a mental image. The priority structure view thus unifies all forms of attention by taking as the primary notion the notion of a mental state's relative priority for the subject. The forms differ only in which aspect of the subject's mind has the highest relative priority. Attention to external objects, on this view, thus gets explained by the relative priority of aspects of the subject mind. The fact that the agent's attention is directed at a cookie in front of her, or at the color of the cookie, is constituted by the fact that the agent is prioritizing a mental state that is about that object or feature. When the agent is prioritizing seeing the cookie, thinking about it, or feeling an urge to eat it, then she is focusing her attention on the cookie in very different ways. Priority orderings thus are more fine-grained than the items the agent is attending to. The priority structure view also takes attention to be graded. A mental state can occupy more or less of the agent's attention depending on where in her current priority ordering it is located.

What we need from the priority structure view, for present purposes, is something fairly minimal and common-sensical: occurrent mental states occupy the agent's attention to various degrees, and the degree to which they occupy the agent's attention is a matter of the priority they have for the agent at the relevant time. It is compatible with this view

that there is some deeper, further account of what it takes for a mental state to have priority to the agent.

In what follows I will argue for two features of these priority orderings: first, for a role that priority plays in coupling an agent's preferences or intentions to her actions. Second, that an agent can intentionally affect the priority ordering of her mental states, based on her preferences for such an ordering, and that she can do so in the relevant self-control situation.

The argument showing how intentional, synchronic self-control is possible, takes the following form:

1 Non-deviant causal links between a preference for A over B and an intentional action A are mediated by the agent's distribution of attention, which is the associated priority ordering of the preference for A. (The Mediation Claim)
2 An agent who prefers A over B can at the same time, psychologically and consistently, prefer not to have the associated priority ordering of the preference for A. In this case, she has a diverging attention preference. (The Attention Preference Claim)
3 If the agent acts on her diverging attention preference, she can intentionally break the causal link between her preference for A, and her A-ing. (The Intentionality Claim)
4 If the agent intentionally breaks the causal link between her preference for A, and her A-ing in this way, she intentionally engages in synchronic self-control. (The Self-control Claim)

So, 5 It is possible for an agent, by intentionally re-distributing her attention, to intentionally engage in synchronic self-control.

Since no willpower, no mental resources, and no mental division are mentioned anywhere in this, this argument shows that it is possible to engage in synchronic self-control without them.

In the next sections, I defend the four premises of this argument.

## 13.6 Preferences and their associated priority structures

Much modelling in decision theory and economics, and much philosophy of action, simply takes an agent's preferences and how they lead to action and choice behavior as given. The mediation claim is the result of thinking more carefully about the link between preference and action.

Let us start with an intuitive idea and compare preferences to desire, where a characteristic link to the agent's distribution of attention has long been recognized: if one has a strong desire for an action, then one's attention will be insistently drawn to appealing properties of that action, or considerations that seem to count in its favor (Scanlon, 1998), and one will be disposed to attend to things one positively associates with what one desires (Sinhababu, 2017). The same is plausible for preferences.

When an agent has a preference for an option, then when it comes to making a decision, her attention will be drawn to something that seems positive or appealing about that option. If Christina, in front of the store, genuinely prefers the cookie over the run, then something motivating about that option must be on her mind at that time. Her attention, whether in perception or thought, must be directed toward the cookie, all that is good about eating it, and other appealing properties of that option. A preference that doesn't direct one's attention seems toothless and without motivational power. For a preference to become alive in an agent's decision making it must at the relevant time engage her attention and prioritize what is relevant for enacting that preference.

In a series of recent publications, Dietrich and List (2013a, 2013b) show how to integrate such intuitive considerations into a formal model of preference and rational choice. On their model, an agent's preference order is given by how she weighs her motivating reasons at the time of decision making. They summarize the model as follows (I won't go into the formal details here):

> …at any time, the agent is in a particular psychological state, represented by his or her set of motivating reasons in relation to the given alternatives, which, jointly with the agent's weighing relation, determines his or her preference order. This preference order then induces a choice function, which encodes how the agent would choose from any concrete set of alternatives.
>
> (Dietrich & List, 2013b, p. 126)

According to Dietrich and List, the agent's preference-based choice of an option thus goes through being motivated by certain properties of that option.

When it comes to what makes a particular property motivating to a specific agent at some specific time, Dietrich and List suggest that at least one way is that the agent focuses attention on those properties when she is forming her preferences or is in a relevant choice situation. Drawing on "the concept of attentional salience as frequently used in psychology and behavioural economics" they suggest that a consideration becomes motivating to the agent if she "focuses on it actively or uses it as a preference-formation heuristic or criterion" (op. cit., p. 109). On the resulting view, then, preferences are formed "by focusing – consciously or otherwise – on certain properties of the alternatives as the motivationally salient properties" (Dietrich & List, 2013a, p. 622), and preferences are changed "when new properties of the alternatives become motivationally salient or previously salient properties cease to be salient" (ibid.).

If an account like Dietrich and List's is right, then preferences are partially constituted by having one's attention drawn or directed to motivating reasons (subjectively motivating properties) for particular actions

or options. Therefore, certain priority structures will be constitutive of those preferences, and all preference-based action will be mediated by an associated priority ordering. Preferences lead to choice behavior in part through activating parts of their associated priority structures.[11]

We can further support the claim that the preference-action link is mediated by attention by drawing on work by Wayne Wu (2014, 2016) about the role of intentions in the production of action. What Wu says about intentions, we will see, transfers to preferences.

An agent's intentions, Wu observes, are not events that happen at a particular time and kick off an action. They are "standing states that persist over time" (Wu, 2016, p. 106). How then do those standing states lead to specific actions at specific times? According to Wu, they are structural causes of action. They structure how an agent selects a behavioral path in a space of behavioral possibilities. In a specific situation, an agent may be aware of various potential targets for action (various things she perceives or thinks of) and in regard to those potential targets, she is aware of various behavioral options: what she can do with the targets. Any agent that does anything (whether intentional or not) must couple the potential targets for action to a behavioral response (Wu calls this the Many-Many problem). The causal role of intentions, Wu argues, is to structure an agent's behavioral space: they bias her choice toward certain types of behavioral responses to certain perceptual situations. Intentions set "the weights that bias which selections are made in action" (Wu, 2016, p. 110).

How is this linked to attention? For Wu, attention just is the selecting of a specific perception-behavior mapping in a specific situation (Wu speaks of attention as selection for action). The causal role of intentions in the production of action, for Wu, is that they bias the agent's attentional system toward selecting certain responses in a certain range of perceptual situations. Intention guides action through the agent's deployment of attention.

The question Wu asks with regard to intention also arises with regard to preferences. An agent's preferences are standing states. An agent has certain preferences over some period of time (on the Dietrich and List model this would be the agent's weighing function). She may have formed those preferences at some particular moment, and she may revise her preferences later. But the preferences themselves are not events that occur in the agent's mind at some particular moment. They are standing states. Given what Christina perceives in her specific situation, or what – more generally – she is aware of in that situation, there are many different things she could do with regard to those things. What role do her preferences play in selecting one behavioral path rather than another? We should be inspired by Wu's account of the role of intentions in the guidance of action: preferences lead to action by biasing the agent's attentional system toward selecting responses in a range of situations.

This Wu-inspired thought again minimally implies that preferences have associated priority structures: a preference is linked to certain ways

of attending. Preferences do not kick off actions by themselves. They cause actions at least in part by adjusting what the agent will attend to in a certain range of circumstances. In my own terminology: the influence of preferences on action is mediated by the agent's priority structures, i.e., by how much attention she pays to what she is aware of in various situations. Preferences set the weights for coupling items in the situation the agent is aware of to behavioral possibilities. They are (at least in part) dispositions to select items the agent is aware of as priorities for action.

We can accept that the influence of preferences on action is mediated through attention without accepting Wu's further claim that attention just is selecting items for action. We can see how by putting together Wu's ideas with those of Dietrich and List: attention mediates between preference and action because a preference disposes an agent to attend to certain items *and their motivationally relevant properties*. It lets the agent see a certain action possibility in a positive light. And seeing it in that positive light is an event that causes the agent to choose the action. Preferences are structural causes of action. Focusing on motivationally relevant properties is an event cause.

We have therefore intuitive considerations for the mediation claim, a formal model that supports it, and the philosophical considerations just mentioned.

The claim that choice behavior is mediated by the agent's distribution of attention is not a mere abstract possibility, though. There is also empirical work in support of it. While non-perceptual forms of attention, according to what I have argued so far, clearly also play a role in linking preferences and behavioral choice, some of the most detailed empirical work has shown the importance also of perceptual attention. Krajbich, Armel, and Rangel (2010), for example, show that a very simple choice between two images is largely predicted by how much time subjects spent looking at the respective images. A large effect remains even when how much a subject *likes* a specific image is controlled for. Many other such effects are known. Based on a review of more than 65 studies, Orquin and Loose (2013) have concluded that "attention plays an active role in constructing decisions" (p. 203). Attention leads to a variety of "downstream effects on decision making" (ibid.). The known empirical work on attention and choice behavior thus like the philosophical consideration supports the view that preference-based choice is largely mediated through the agent's distribution of attention leading up to and at the time of decision making.[12]

## 13.7 Action preferences and attention preferences

If the link between preference and action is mediated by associated priority structures, as I have argued, then it is possible to change whether the action results by interfering with the priority structures. In this

section, I will argue that the agent can *intentionally* interfere with those priority structures and thereby intentionally decouple the action from her preferences. I will again focus on philosophical considerations for this claim, but I take those broadly in line with the conclusion drawn by Orquin and Loose in the review just mentioned, that the role of attention in decision making is "constructive" since decisions emerge "not as a simple application of preferences and heuristics to choice stimuli but, through complex interactions among stimuli, attention processes, working memory, and preferences" (2013, p. 203).

An agent's preference for an option disposes her to pay more attention to motivationally relevant properties of that option as well as aspects of what she is aware of that will couple her preference to a situation specific action that enacts that preference. This is the associated priority structure of that preference. The point of this section is to argue that this does not imply that the agent also *prefers* to attend to those motivationally relevant properties and perceptually presented items. I will thus argue for the attention preference claim, i.e., the claim that an agent who prefers A over B can at the same time, psychologically and consistently, prefer not to have the associated priority ordering of the preference for A, and that she can intentionally act on that diverging attention preference.

For this discussion it is important to recognize that attention is a form of agency just like embodied action (cf., Chapter 7 of (Watzl, 2017, pp. 138–155)). An agent can choose to focus her visual attention on an item just like she can choose to pick up that item with her hand. And she can control the focus of her attention in ways that are highly similar to the way she controls her bodily movements. An agent who, for example, engages in a visual search task (looking for Waldo in the famous cartoon drawings) may explicitly adopt a particular strategy: she may scan the picture with her visual attention from left to right and top to bottom.

Specifically, what we need here is the claim that agents often have preferences regarding what to attend to. When listening to a piece of orchestral music, someone with an interest in the flute, might prefer to focus her auditory attention on the melody of the flute over focusing it on the sound of the violins. And you might prefer to look at one of the photographs in your office over some of the other photographs. Agents also often have preferences regarding which *features* to attend to: in one situation, the agent might prefer to focus on the rhythm of the melody that the flute plays rather than how loudly the notes are being played, while in a different situation, she might prefer to focus on the loudness instead. Or she might prefer to look at the shape of her dining room table (a very pleasant shape) rather than visually focusing on its color (rather less pleasant). The same holds for attention in thought: when thinking about his last vacation, Amira might prefer focusing his attention on how peaceful it felt rather than on the fact that it was raining most of the time. Or consider attention to bodily sensations: you might have a

preference for attending to the aftertaste of coffee in your mouth over attending to the unpleasant numbness in your leg.

In other words, agents have preferences over priority structures. They prefer to have some priority structures over others. These preferences, of course, are often not extremely fine-grained: there are many possible priority structures that are compatible with an agent's preference for focusing most of her attention on the rhythm of the flute. The attention preference need not specify how much attention she pays to the violins and how much should be left for visually taking in her surroundings. The same holds for other preferences: Christina's preference for eating a cookie over going for a run does not specify which cookie she eats, the details of how she is going to get it, or how she will put it in her mouth. I will thus assume that agents have attention preferences: ways of attending (i.e., priority structures) are among the options agents' preference relations range over.

One might ask: how is the idea that we have attention preferences compatible with the claim of the last section that attention mediates between preference and action? Isn't there some problematic circularity or infinite regress? No. Attention also mediates between an attention preference and the act of attention. Suppose you prefer paying attention to painting A over paying attention to painting B. How does that preference actually get you to pay attention to A rather than B? The mediation claim applies here just like it applies in other cases. Your preference disposes you to focus on properties that make paying attention to A appealing, and it structures your attention at one time to select targets for your attention at a later time. Once the motivating properties of paying attention to A are salient to you, you likely end up paying attention to A. There is no circularity or infinite regress. A preference for one priority structure over another biases an agent's *other*, *mediating* priority structures toward positive features of that priority structure. What would lead to an infinite regress is if I were to claim that the only way an agent could arrive at the *mediating* priority structures is again via a preference for those mediating structures. But this is not what is claimed here. The problematic claim would be analogous to claiming that the only way an agent's preference for an overt action could lead to that action is if the agent also had a preference for everything that mediates between her preference and that action. But such a claim has no plausibility.

Now suppose that an agent has a preference for some overt action A, like eating a cookie, over another overt action B, like going on a run. That preference has an associated priority structure. Call that $P_A$. Does this entail that the agent at that time also prefers to have that priority structure over relevant alternatives, like the priority structure associated with B, $P_B$? No. The fact that $P_A$ mediates between the agent's preference for A and A does not entail that the agent also prefers $P_A$. We have, in fact, already seen this in the last paragraph: a priority structure might

mediate between a preference and an action without an agent also preferring to have that priority structure. This shows that it is possible that an agent prefers option A over an incompatible option B, yet she also prefers not to have the priority ordering associated with A, but rather have a priority ordering associated or at least compatible with B. In other words, it is possible that the agent's attention preference diverges from her action preference.

One might ask: even if diverging attention preferences are possible in principle, is such divergence ever psychologically realized? I think it clearly is. Suppose you are at a supermarket, preferring to eat a cookie, which draws your attention to various motivationally salient properties of that action like the location of the cookies in a supermarket shelf. But you need not *prefer* to focus your attention on that location in the shelf: you want to eat the cookie, but I also want to buy some laundry detergent (though you want the cookie more). You might prefer to focus your attention on where the detergent might be, since you haven't found it yet. Your attention is *drawn* to the cookie locations, but you prefer it to be somewhere else. This is a perfectly ordinary situation, even aside from self-control cases.

Do diverging attention preferences necessarily make the agent's preferences inconsistent or incoherent? I don't think so. An agent might consistently and coherently prefer A over B, and C might be necessary for A, and yet the agent might prefer not to be in C. If the agent can be in C unintentionally (i.e., not on the basis of a preference for being in C), then the agent need not prefer C over not-C in order to get to A. Maybe a coherent agent must prefer to take the means if she prefers the end. If that is true, then it is true only because means are intentional actions the agent must perform in order to get what she wants. But priority structures are not a means an agent takes toward an end she aims to achieve. Priority structures mediate between preferences and action, but not as intentional means to the action. Consider this analogy: in order to pick up a coin, I must move my fingers in a specific way. I can move my fingers in this specific way intentionally, but in order to pick up the coin, I need not, and will normally not, move them intentionally in this way. Therefore, there is no incoherence in assuming that while I really want to pick up the coin, I prefer not to move my fingers in the way that is required.

Still, one might say, there is surely some kind of conflict when one has an attention preference that divergences from one's preferences for overt action. And yes: of course, there is. If the agent acts on her attention preference, she will not realize her overt action preference. And if she acts on her overt action preference, she will not realize her attention preference. There is a subjective conflict. The agent can't get everything she wants. That just *is* the subjective conflict that, among other situations, characterizes self-control situations. The existence of a subjective

conflict when the agent has divergent attention preferences is not a problem for the present view. It is part of the datum we wanted to explain.

I thus hope to have convinced you of the attention preference claim. It is psychologically and consistently possible for agents to have attention preferences that diverge from their action preferences.

## 13.8 Intentionally breaking the preference-action link

I still need to defend the last two premises of the argument. In this section, I defend the intentionality claim, i.e., that if the agent acts on her diverging attention preference, she can intentionally break the causal link between her preference for A, and her A-ing.

Preferences are structural causes of action, I have argued. They bias the agent's attention system toward prioritizing action relevant objects and their appealing properties. Such appealing properties will be salient to the agent: her attention tends to get drawn to them. But if the agent has a diverging attention preference, she prefers to have a different priority structure instead. If she acts on that attention preference, she shifts her attention in a different direction or suppresses the relevant saliences. Since the causal link between her action preference and her action is mediated by those saliences, interfering on the mediating structure will unlink the preference from the action. Since that interference is based in the agent's preferences, and there are no deviant causal chains, the interference is intentional. Therefore, if the agent acts on her diverging attention preference, she can intentionally break the causal link between her original preference and her action. Which is just the intentionality claim.

Let me illustrate how agents intentionally unlink their preferences from action with one of the most famous experimental studies of self-control: the delay of gratification experiments, pioneered by Walter Mischel in the 1960s.[13] Here children are presented with two options. The first option is to eat one cookie (or marshmallow) now. The second option is to wait and get *two* cookies later. The cookie in front of the child, we assume, has a powerful grip on her, while the child views the second option as subjectively better (two are better than one). Once the child is sitting in front of the one cookie and the experimenter has left the room, she prefers the one cookie over waiting for a second cookie later. But she thinks that she should wait for the second cookie. One nice feature of the delay of gratification paradigm is that it gives us a quantifiable measure of a self-control strategy's effectiveness in terms of the children's waiting time.

The delay of gratification results illustrate how preference is mediated by attention, and how this allows for intentional interference by the agent.

One important result is that brute effort of attention (unlike what William James may have thought) is not an effective strategy. Suppose

that the experimenter primes subjects to try hard to keep attention focused on the two-cookie goal by, for example, instructing children to mentally rehearse "I will get two cookies later". Keeping the two-cookie goal mentally alive prioritizes mental images of or thoughts about two cookies. Such strategies shorten waiting times relative to a neutral base line. With the mediation claim, we can easily see why: thinking of two cookies is thinking of cookies. Activating the goal of two cookies makes the one cookie in front of the child, and especially its yumminess, extremely *salient* to her. The child's attention is drawn to a motivationally salient property of the one cookie. Activating the two-cookie goal thus supports the priority structure that mediates between the preference and the cookie. Therefore, it makes the child more rather than less likely to fall for the temptation. Preferring to think of two cookies later is not, after all, a diverging attention preference.

Effective strategies, by contrast, involve intentionally re-orienting attention in a way that *lowers* the priority ranking of states that mediate between the child's preference and the action of eating cookies. One effective strategy is self-distraction. The child might think about a fun activity or sing songs in her head. These are intentional actions. Children who are primed by being informed of the availability of such strategies, and thus are more likely to use them, have increased waiting times. Since here the children are intentionally prioritizing images and thoughts that have nothing to do with cookies, these strategies, unlike refocusing on the two-cookie commitment, do not make the one cookie more salient: by increasing the salience of, say, mental images of playing on the beach, the relative salience of the yumminess of the cookie is decreased instead. Similarly, waiting times are increased when the child intentionally focuses attention on shape or color of the cookie. Now she is deprioritizing the perception of the affective properties (its yumminess) of the cookie by prioritizing its non-affective properties (its color or shape): she intentionally shifts attention away from motivationally salient properties of cookie eating.

In order to make something intrinsically boring, like the shape of a cookie, easier to attend to, techniques for reconstrual help, like thinking of the cookie as a UFO and of the raisins as the aliens that ride it. It also helps to build associations between seeing the tempting object and something else (you may train yourself so that every time you see a cookie you think about salads). The formation of such habits does not make the relevant shifts of attention automatic or unintentional. A habit-based action like switching on the light when entering your house or putting on your running shoes when the clock strikes, is still an intentional action. Implementation intentions (Gollwitzer & Brandstätter, 1997) are still intentions. And habitual preferences are still preferences.

Of course, a child, or an adult, must know what to attend to in order to live by a specific normative judgment about what she should do. By

acting on a diverging attention preference it is *possible* to intentionally disconnect a preference from an action. But this does not entail that you succeed. Self-control attempts by means of intentional re-prioritization might fail, and they often do. The relevant attention preferences must be linked up with the subject's normative judgment and they must be effective: as the case of focusing on the two-cookie goal shows, what the agent *thinks* is a priority structure that will decouple her preference from action might not actually be a priority structure that has this effect.

## 13.9 Attention preferences and the judgments that drive them

Suppose that the agent intentionally breaks the causal link between her preference for A, and her A-ing by a shift of attention that destroys the associated priority structure of that preference. Does that amount to an intentional act of synchronous self-control? According to the last step in my argument, the self-control claim, it does. I argue that it does by answering two objections. One is that the relevant avoidance of the loss of control threat isn't really synchronous. The other that while it might be an intentional action, it is not an intentional act of self-control.

With regard to synchronicity, the objection would be that the re-prioritization strategy really describes diachronic self-control. Specifically, let us ask whether the account can explain *full-blooded* self-control, which Sripada (2014) raises as a problem for a view of self-control by Alfred Mele (1992) with which the present account has some similarities. (For the objection that Mele's view only allows for diachronic self-control see also Kennett and Smith (1997).)

According to Mele, agents engage in synchronic self-control by engaging in a second, *ancillary*, intentional action that is compatible with acting on her wayward preferences, and that will, in due course, change her motivation. The re-prioritization account agrees that self-control is achieved through a secondary action. Here this is an intentional change to the agent's priority structures.

Full-blooded exercises of self-control, according to Sripada, are cases where the agent never even begins to act on her wayward preference (or as he says, her "strongest desire"). Sripada believes that there are such cases. Christina might never even begin to take a step toward the pastry shop, and Amira might not begin to initiate his punch. Sripada objects that Mele's view, by contrast, entails that the agent will always begin to engage in the preferred and wayward act. On Mele's view, the agent begins to engage in the wayward action; while it is on its way she engages in a compatible ancillary action, like telling herself "Remember the run!". This, Mele plausibly believes (and Sripada and Kennett, and Smith agree), may change her motivations (on Mele's view this indeed often happens through a shift in attention), which in turn interrupts her

tempting action, and thus the agent achieves self-control. Sripada thinks that this account neglects the full-blooded cases.

Does the re-prioritization account, arguably like Mele's, rule out such full-blooded self-control? No. An agent's preferences, as I said, always lead to action through her priority structures. Since re-prioritization acts on these priority structures, the agent may never initiate the wayward action. Christina and Amira can decouple their preferences from action through a shift in priorities, before they initiate their preferred action. The re-prioritization account differs from Mele's in exactly the crucial respect: since priority structure mediate between the preference and the action, the act of changing those structures is not one the agent performs while already engaged in her preferred action; she performs it before she ever engages in it.

Sripada and Kennett and Smith might object that the agent *does* initiate her preferred wayward action. Her cookie preference makes appealing properties of eating cookies salient to Christina, and draws her attention to where the cookies are found. Even if she interrupts herself by, say, shifting her attention to a mental image of her last vacation, she has already begun to act on her preference. The crucial question seems to be: is prioritizing positive features of eating cookies already a part of the action of eating cookies, or is it not? If we think of actions as the event specified in the content of the agent's preferences, then the associated priority structures are not part of the action, and the present objection fails. If, by contrast, we think of actions as complex temporally extended processes that include the mental precursors of the event specified in the agent's preferences (cf. Dretske, 1991), then the associated priority structures are part of the action. But on this way of individuating action, the agent's preferences (and her beliefs) are *also* part of the action: the action is the process of the preferences and beliefs leading to the bodily behavior. But then the action has already begun when Christina first forms her preference for that cookie, and full-blooded self-control would be ruled out by the very description of the loss of control threat situation. Sripada and Kennett and Smith therefore cannot appeal to it.

I, therefore, conclude that the re-prioritizing strategy makes self-control synchronic in the relevant sense.

The next objection is that while the agent may do something intentionally to avert the loss of control threat (based on her attention preferences), what she does intentionally is not an intentional act of self-control. The agent must engage in the act *in order* to do as she believes she should do. The mere fact that she does something that as a side-effect leads the agent to choose the option she judges to be correct is not enough to show that her act was intentional *qua* self-control act.

I agree that an intentional self-control act must be the result of the agent's view of what is the subjectively better option, in our case, her normative judgment. If a child in Mischel's experiment just happens to

prefer focusing her attention on the shape rather than the yumminess of a cookie, and as a result waits longer, this would not be an intentional act of self-control. We have to demand, indeed, that the agent must form her relevant attention preference based on the reasons for the better option she is aware of in her normative judgment. But it is guaranteed that there are such reasons the agent is aware of: in a subjective conflict situation that characterizes self-control cases the agent is aware of the better option as better. So, she sees reasons for choosing the better option. For an intentional self-control act, those reasons must inform the formation of her attention preference by means of which she decouples her momentary preference from action. Those reasons clearly are motivationally relevant for what to attend to in that situation. Thus, by the Dietrich and List (2013a, 2013b) account, she can form that attention preference by focusing on those reasons.

Of course, as we have already seen, the fact that the agent's attention preferences are informed by her reasons in this way does not mean that she knows how to attend in order to secure the better choice. The child might try to focus on a mental image of two cookies in order to resist her temptation. But that, as we have seen, is not an efficient self-control strategy. It is still an intentional self-control attempt, since the child might re-focus her attention in this way based on the reasons for the better option she is aware of. And the attempt might, in some cases, and for some period of time, succeed. In this case, it would be a synchronic act of self-control. For more efficient self-control, the agent must have, what might be called, an *attention skill*: she must be disposed to bring to bear situation specific knowledge of what to attend to in order to secure her the choice of the better option (cf. Pavese, 2016). Having a disposition to bring to bear certain knowledge in the relevant situations thus is part of how we succeed in self-control (here I agree with ideas in (Kennett & Smith, 1996, 1997)). But when we re-focus our attention based on that knowledge we engage in an ordinary intentional action, based on our preferences, which we form in light of that knowledge. Synchronic self-control acts can be intentional attempts *to* secure the better option.

## 13.10 How to live without special motivational powers

I hope to have shown that we can explain synchronic self-control without appeal to any special motivational powers. Once we understand how the horses pull the chariot, i.e., how preferences are linked to action, we need no more than horses. I also hope to have shown that the possibility of such self-control follows naturally from the role of attention in the control of all action. The resulting picture, I believe, is psychologically plausible as a description of how agents actually engage in self-control and may succeed at it.

I will end by answering two further questions.

First, according to the re-prioritization model, self-control involves no willpower, and no battle between two motivational systems. What then explains the phenomenology of effort that seems to accompany self-control? If self-control just is an ordinary action, why does it feel so hard?

In my view, the sense of effort is probably the result of a number of features of self-control situations.[14] First, re-prioritization in order to achieve the subjectively better result is an error-prone process with an uncertain outcome. The agent cannot be sure that the way she chooses to re-focus attention will actually lead to the correct result, and it is easy to make mistakes. It has been shown that experienced effort and judgments of error-likelihood are strongly correlated (see Inzlicht, Schmeichel, and Macrae (2014) for this idea, and Dunn, Inzlicht, and Risko (2019) for some detailed results). Second, the agent's attention will likely fluctuate between, on the one hand, priority structures that facilitate choice of the preferred option and ones that accord with the one she judges to be better: given that the agent's action preferences persist, they will keep drawing her attention to relevant items and their appealing features even though the agent prefers to have a different distribution of attention. Such fluctuations of attention between conflicting cues and the associated valuation fluctuation have been shown to be linked to experienced conflict and task difficulty (Kiani, Corthell, & Shadlen, 2014; Krajbich et al., 2010; Desender, Van Obstal, & Van der Bussche, 2017; cf. Berkman et al., 2017b). Third, there is the hypothesis that subjective effort is linked to opportunity costs (Kurzban, Duckworth, Kable, & Myers, 2013): given that the subject is aware of two options, one of which she prefers and the other she judges to be best, effort might signal that there is another alternative distribution of attention that the agent is not having. Generally, given the complexities of the psychological processes known to be involved in the generation of subjective effort, it would be a mistake to take such subjective effort as much evidence for the existence of a willpower faculty or a battle between two motivational systems. The re-prioritization account is consistent with what is known about the experience of effort.

Second, and relatedly, one might ask how – on the present view – we are supposed to explain why self-control attempts seem to show ego-depletion effects of the kind summarized in Baumeister et al. (2018). Holton (2009), for example, mentions those findings in support of the view that, like a muscle, willpower may become weak, and can be trained by its repeated use (appealing to Muraven and Baumeister (2000)).

Self-control can be trained because attention skills can be trained. It is thus compatible with the re-prioritization account that self-control can be improved through training. What the account denies is that willpower can be trained *like a single, specialized muscle*. And for that, the evidence is decidedly mixed (cf. Inzlicht et al., 2014). Generally, the

evidence for ego-depletion has come under attack and is subject to replication failure (Carter and McCullough, 2014). One of the most surprising apparent findings seeming to support the existence of a psychic resource was that this resource was allegedly measurable in terms of blood glucose levels (Gailliot et al., 2007). Yet, it isn't clear that the results that self-control decreases glucose levels hold water: even rinsing the mouth with glucose might enhance self-control (Sanders, Shirk, Burgin, & Martin, 2012). Based on such findings, it has been argued that we can better explain the effects of glucose on self-control without positing any psychic resources. Kurzban et al. (2013), for example, argue that the detection of glucose signals "success" and is thereby motivating (see also, Levy, 2016). Generally, I agree with Inzlicht et al.'s (2014) assessment that "mental resources" may be, just as David Navon (1984) argued more than 30 years ago, a "theoretical soup stone", a mysterious entity that seems important, but that in fact is explanatorily empty once we have a fuller picture of the mechanisms actually involved in self-control.

## 13.11 Conclusion

I have argued that we can substantiate the intuitive idea that attention and self-control are connected. Self-control is indeed a matter of shifting the focus of attention in the right way. Yet, self-control is mostly not achieved by an effort of attention, trying hard to keep one's better options in clear view. Rather, it is achieved through a complex set of attentional skills. Attention is employed in self-control just like it is employed in other forms of agency. It acts as a flexible interface between our standing preferences and our actions. Agents have a form of agential flexibility over action because they need not translate their preferences directly into action. Through attention, agents gain a form of freedom: that freedom, though, is not a special freedom that shows the exercise of a mental faculty or an aspect of the mind that comes into play only when self-control is called for. Rather, it is a freedom that attention weaves into the structure of agency quite generally.

My view of self-control is thus deflationary: we use attention in order to realize our goals, short or long term, we use it to act on our commitments or in favor of situational demands, we use it to realize what we think is right, or we use it to act against our best judgment. While some have argued that there is a special rational demand to be enkratic (cf. Broome, 2013) – whether or not that is ever achieved through special powers – I believe that the deflationary view argued for here lends some further support to views on which there is no such demand (cf. Audi, 1990; Arpaly, 2004; Brunero, 2013; Reisner, 2013). Christina's judgment that she should run and eat salad might be the result of a skewed deliberative process, channelled by problematic societal pressures

regarding ideals of beauty and health. The cookie might in fact be good for Christina, making her a better runner, and a happier person, and her "temptation" in front of the store might be the result of sensitivity to exactly those reasons. Amira's anger, similarly, might be a rational response to the policemen's behavior, making Amira more rather than less sensitive to the normative reasons for action present in the situation (see Srinivasan (2018) on anger and D'Cruz (2013) on reasons that might in principle be inaccessible through deliberation). Self-control might be irrational in the sense that it makes Amira less responsive to the reasons present in the situation at hand. If self-control is not achieved by reason winning over passion, or by the self taking control of the horses, then plausibly self-control also need not always be what is most rational. If the normative structure of the situation is anything like the descriptive structure, then practical wisdom is likely "unprincipled" (Arpaly, 2004).

## Notes

1. Plato, Phaedrus, 246a–254e. See also Ganeri (2018) on the imagery in Upaniṣads written around the same time. As Ganeri points out, the Buddhists empathetically denied the adequacy of the image.
2. Sinhababu (2017) argues *for* Humeanism and against, for example, willpower (Ch. 8) and a substantial role of the self (Ch. 10), by appealing centrally to the role of attention in the link between desire and action. (He suggests also that certain patterns of attention are constitutive of desire; cf. Ch. 1 and Ch. 5). Yet, if attention cannot be successfully reduced to a combination of representational (beliefs) and motivational states (desires) (cf. Watzl, 2017), and if attention plays an irreducible role in the (rational) explanation of action (which I argue for below, but see also Wu (2014; 2016)), then Humeanism is false in an important sense.
3. Cf. Brownstein (2018).
4. Thanks to Johannes Rössler, who pushed me to say more about how my view treats cases like the Angry Punch.
5. It has not escaped my notice that both of my examples have problematic features. I will return to the issue of normativity and what is good for the agent in the conclusion.
6. The non-agential view of self-control (cf. Kennett op. cit.), in my view, need not deny this datum. In Section 13.4, I argue that this view fails, but not necessarily because it cannot explain that self-control is non-accidental.
7. The capacity for self-control is at issue when one argues that infants or children have less self-control than adults, or that self-control develops throughout childhood, puberty, and early adulthood. Tangney, Baumeister, & Boone (2004) have developed a "self-control scale" aimed at measuring capacity for self-control. For some problems with such a scale see, e.g., Brownstein (2018). The deflationary view developed in this paper may be used to further cast doubt on the use of such psychological measures.
8. Thanks to Katharine Browne and Jurgis Karpus for pressing this point on me.
9. Okasha (2016) argues that behaviorism about preferences might be correct for a normative (and not descriptive) theory of rational choice. This, though, is not what we are concerned with at this point.

10. Cf. Inzlicht, Schmeichel, & Macrae (2014); Inzlicht & Berkman (2015); Berkman et al (2017b). For a related discussion of the role of attention for various forms of control in the literature on artificial intelligence, see also Bello and Bridewell (2017).
11. Note that in order to accept this last claim, we don't need to define preferences in terms of those associated priority structures here.
12. A particularly powerful illustration of the interaction of attention and choice concerns drug addiction. It has been found that drug-related features are highly salient for the addicted person and consequently tend to draw her attention away from affectively more neutral features. This has been tested through the so-called addiction Stroop test, where subjects must report on the color in which a word is written. Here addicted people show longer reaction times and higher rates of mistakes when asked to report the color of drug-related words compared to the control group. For the addict "salience attribution transforms the sensory features of the incentive stimulus into an especially salient percept, which 'grabs attention', becomes attractive and 'wanted' and thus guides behavior to the incentive" (Robinson & Berridge, 1993, p. 261).
13. See Mischel, Ebbesen, and Raskoff Zeiss (1972); See Mischel (2014) for a popular overview.
14. In a related context, Sinhababu (2017) suggests that we should explain the sense of effort involved in self-control in terms of effort of attention. This, though, clearly begs the question (what, after all, would explain the sense of effort of attention?).

## References

Arpaly, N. (2004). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Baumeister, R. F., Bratslavsky, E., & Muraven, M. (2018). Ego depletion: Is the active self a limited resource? In Baumeister. In R.F (Ed.), *Self-regulation and self-control* (pp. 24–52). New York: Routledge.

Bello, P., & Bridewell, W. (2017). There is no agency without attention. *AI Magazine*, *38*(4), 27–34.

Berkman, E. T., Livingston, J. L., & Kahn, L. E. (2017a). Finding the "self" in self-regulation: The identity-value model. *Psychological Inquiry*, *28*(2–3), 77–98.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017b). Self-control as value-based choice. *Current Directions in Psychological Science*, *26*, 422–428.

Bermúdez, J. L. (2009). *Decision theory and rationality*. Oxford: Oxford University Press.

Broome, J. (2013). *Rationality through reasoning*. Oxford: John Wiley & Sons.

Brownstein, M. (2018). Self-control and overcontrol: Conceptual, ethical, and ideological issues in positive psychology. *Review of Philosophy and Psychology*, *9*(3), 585–606.

Brunero, J. (2013). Rational akrasia. *Organon F*, *20*(4), 546–566.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823.

D'Cruz, J. (2013). Volatile reasons. *Australasian Journal of Philosophy*, *91*(1), 31–40.

Desender, K., Van Opstal, F., & Van den Bussche, E. (2017). Subjective experience of difficulty depends on multiple cues. *Scientific Reports*, *7*, article number 44222, 1–14.

Dietrich, F., & List, C. (2013a). A reason-based theory of rational choice. *Nous*, *47*(1), 104–134.

Dietrich, F., & List, C. (2013b). Where do preferences come from? *International Journal of Game Theory*, *42*(3), 613–637.

Dietrich, F., & List, C. (2016). Reason-based choice and context-dependence: An explanatory framework. *Economics & Philosophy*, *32*(2), 175–229.

Dretske, F. I. (1991). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35–55.

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: Roles of perceived error-likelihood and time demands. *Psychological Research*, *83*, 1033–1056.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M. … Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*(2), 325–336.

Ganeri, J. (2017). *Attention, not self*. Oxford: Oxford University Press.

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology*, *73*(1), 186–199.

Hausman, D. M. (2000). Revealed preference, belief, and game theory. *Economics & Philosophy*, *16*(1), 99–115.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford: Oxford University Press.

Inzlicht, M., & Berkman, E. (2015). Six questions for the resource model of control (and some answers). *Social and Personality Psychology Compass*, *9*(10), 511–524.

Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127–133.

James, W. (1890 [1981]). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Kennett, J. (2001). *Agency and responsibility: A common-sense moral psychology*. Oxford: Clarendon Press.

Kennett, J., & Smith, M. (1996). Frog and toad lose control. *Analysis*, *56*(2), 63–73.

Kennett, J., & Smith, M. (1997). Synchronic self-control is always non-actional. *Analysis*, *57*(2), 123–131.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

Levy, N. (2016). The sweetness of surrender: Glucose enhances self-control by signaling environmental richness. *Philosophical Psychology*, *29*(6), 813–825.
Mele, A. R. (1992). *Springs Of action: Understanding intentional behavior*. Oxford: Oxford University Press.
Mischel, W. (2014). *The marshmallow test: Mastering self-control*. New York: Little, Brown and Company.
Mischel, W., Ebbesen, E. B., & Raskoff Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*(2), 204–218.
Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*(2), 247–259.
Navon, D. (1984). Resources – A theoretical soup stone? *Psychological Review*, *91*(2), 216–234.
Okasha, S. (2016). On the interpretation of decision theory. *Economics & Philosophy*, *32*(3), 409–433.
Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*(1), 190–206.
Pavese, C. (2016). Skill in epistemology i: Skill and knowledge. *Philosophy Compass*, *11*(11), 642–649.
Reisner, A. (2013). Is the enkratic principle a requirement of rationality? *Organon F*, *20*(4), 437–463.
Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247–291.
Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.
Sanders, M. A., Shirk, S. D., Burgin, C. J., & Martin, L. L. (2012). The gargle effect: Rinsing the mouth with glucose enhances self-control. *Psychological Science*, *23*(12), 1470–1472.
Scanlon, T. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.
Schellenberg, S. (2019). Perceptual consciousness as a mental activity. *Noûs*, *53*(1), 114–133.
Sinhababu, N. (2017). *Humean nature: How desire explains action, thought, and feeling*. Oxford: Oxford University Press.
Srinivasan, A. (2018). The aptness of anger. *Journal of Political Philosophy*, *26*(2), 123–144.
Sripada, C. S. (2014). How is willpower possible? The puzzle of synchronic self-control and the divided mind. *Noûs*, *48*(1), 41–74.
Tangney, J., Baumeister, R., & Boone, A. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324.
Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford: Oxford University Press.
Wu, W. (2011). Confronting many-many problems: Attention and agentive control. *Noûs*, *45*(1), 50–76.
Wu, W. (2014). *Attention*. Oxford: Routledge.
Wu, W. (2016). Experts and deviants: The story of agentive control. *Philosophy and Phenomenological Research*, *93*(1), 101–126.

# Index


accomplishment process 145
achievements, inference and 128
act of choosing 229
acting on oneself: making inference 129–132; way to 129
action: as act of will 124; capacity 61, 68–71; -driven approach 221; and mere behavior 19; preferences 286–290
action explanation: and interventionism 164; selection criteria, enforcing 171; *see also* naive action explanation
action theory of attention 1–3; to bring inferential power to interventionist models 165; general phenomenon in 46; interventionist models and 165; selection for 68, 83; standard approaches in 148–149
active powers 230
active reasoning 143
active remembering 150
addict, unwilling 14–15
addiction Stroop test 298n12
adverbialism, attentional 81–83
agency: attention and 273; core of 19; digger wasp 22–23; honeybee 23–24; ingredients of 21; level of 24; nature of 19–25, 63; notion of movement 20–22; paramecium 20–21; primitive 20–21; terms of control 156; Velleman's proposal 17
agency theory 234; conceptual limitation of 70; perspective of 73
agential capacity 7
agential explanations *see* explanation
agentive complexity 15; Brent's proposal 18–19; Velleman's proposal 15–16; Wallace's proposal 16–17
agents 123–126; into action explanation 19; appearing 25–26; -causalism 233–234; disappearing 14–19; non-agential systems 19; planning 24–25; psychological 25
Allport, A. 65, 66, 81–82, 84, 89
alpha isoform of calmodulin kinase II (α-CaMKII) 213
Alvarez, M. 123, 247
Andersen, H. 5–7, 164–180, 219–220
Anscombe, G. E. M. 179
appearing agent 25–26
Aristotle 228, 235, 236–237
Armel, C. 286
artificial intelligence (AI) 214; Frame Problem for 214; Good Old-Fashioned 214
attending: as mental action 71–75; mental action as 75–76
attention 79; agent's distribution of 279; biological reality of 64, 67, 73–74; central role as guiding or informing response 67; deduction and mental action 148–153; defined 61; description of 65; determination, entailment to 90–95; effort of 272; as emerging from solving Many-Many Problem 65; intellectual 151; manifestations of 81; by mental action, entailment of 87–90; metaphysics of 67–68; most general mental act 87–95; nature of 81; and non-agential view 279–280; preferences 286–290; psychological reality of 62, 65–68, 71; re-focusing

of 279; role of 273, 280–281; selection for action theory 65–68; significance of 272; ubiquity and heterogeneity of 80–87; *see also* action theory of attention; covert attention; covert perceptual attention; perceptual attention; ubiquity of attention
attentional adverbialism 87; diagnosis for Allport's suspicion 82; objection to 83; overreacting to heterogeneity 81–83; selection for action view and 94
attentional capture 74
attitudes and action 199–200
Audi, R. 128–129
auditory covert attention 71–72
automaticity and control, analysis of 64–65

Bach, K. 123
Baddeley, A. 214
ballistic process: account of conscious mental action 106–107; account of mental action 102–106
Barrett, L. 220
Baumeister, R. F. 295
behavior space: action capacities and 68–71; agent's acting 62; apparatus of 67; attending as mental action 72, 75; concept of 62; conception of action 67; constituting 62–63; depicting production of action 68; framework of 62, 64; human agent 63; mental action as attending 75–76
behaviorism 212, 297n9
beliefs/believing 110, 111, 129; and action 182, 184–185; changing or sustaining state of 130; and choices 182, 187; detective 256; evidence and 134; ideas about 130; inference 122–138, 151, 155; and intentions 155; maintaining system of 130; model of agency in terms of 273; objects of 131; onset of 126–129; and owning 131; premise 147–148; responsible for 122; scientific reason to 259; second-order 146; self-attribution of 55; something for reasons 144; standing 154; suspending 134; tendency to hypostasize 130
Berkman, E. 273
biasing, intention as 64–65
Bickle, J. 213
biological systems, analogous issue in 219–220
Block, N. 205
bodily action 35–36, 89–90, 103, 106; attention in 75; attentive performance of 90; focusing on 64; intentional 107; issue of dividing mental from 70; non-bodily action 70
Boghossian, P. 52, 129, 142–143, 145
'bottom-up' attention 93
Boyle, M. 6, 8–9, 130–131, 182–201
Boyle's law 216
Bradley, F. H. 82
Bratman, M. 24
Bratslavsky, E. 295
Brent, M. 1–11, 18, 100–114
Broome, J. 129, 143
Buckareff, A. A. 6, 7, 228–249
bundle theory of substance 238
Burge, T. 20–21, 24

Cameron, R. 241
Camp, E. 23–24
capacity for action *see* action capacity
Carroll, L. 52, 146–147
Carruthers, P. 185–188
Cartwright, N. 172
Cathy Hutchinson 105
Causal Bayes Nets modeling 165, 171
Causal Exclusion problem 179, 238
causal explanation 215–216; appeal to mental contents in 212; degree of strength of connection 170; generation of 170; involving mental actions 19; of mental events 218; and naive action explanation 167–168; prevalence of commonsense 211
Causal Faithfulness 219
causal Markov and faithfulness: assumptions 165; Causal Faithfulness condition 172–173; Causal Markov condition 172; failure of 172; justifying inferences in 171–173
causal modeling: causal Markov and faithfulness justifying inferences in 171–173; and efficacy of action 164–180; model versus system as primary target of inquiry 168–171; naive action explanation

and rationalization 166–168; Rationalization condition 174–179
causal structural equation modeling 165
causation 123–126; and action 179; by actions 234; agent 123–126; attention in 68; interventionist approaches to 164; of judgment and choice 195; manipulability and 215; mental 238; mere 165, 174; metaphysics of 108–109; non-deviant 15; non-reductive agent 18; rationalization and 179–180; substance 18; transitivity of 52
Charles, D. 236
choice 132–138
Chrisman, M. 130
Cisek, P. 221
Clarke, R. 265
Classical approach 216, 221
Classical Cognition: failure of 210; hypothesis 209; Semantic Efficacy and 208–211
cocktail party effect 71–72, 93, 94
cognitive capacities 104; activity of 107; causal role in terms of 102; manifesting of 109, 111; pertinent 105; problem of understanding 222–223; properties 109; relevant 104–106, 110–111, 113
cognitive decisions 255
cognitive resources 81–82, 85, 152
cognitive sciences: Computational explanations in 209; mental contents in 213–214; psychology and 212; Semantic Externalism in 207; theories in 214; *see also* Classical Cognition; Embodied Cognition; Semantic Efficacy
cognitive system 3
cognitive unison 82, 83, 94, 95n5
cognitive "virtual machine" 221
Collingwood, R. G. 182–183, 198–199
complex contents 55
complex mental action 31
computational explanations in cognitive sciences 209–210
computational theory 68
computationalism 208, 209
conflict resolution 275–276
conjunctions of powers 247
conscious agent 112–113, 116n28; causal power 109; causal role as 108, 111; performing mental actions 106, 107; practices requiring 112
conscious mental action: absence of 113; alternative account of 101, 112–113; ballistic account of 106–111; limiting account of 103; performing 102; salient during 111; skeptical account of 102; standard accounts and 100–101; temporally-extended 112
consciousness 260; agent's stream of 152; ballistic process 102–106; of decision 259; delivering content to consciousness by 111; intentionally delivered to 101; involuntary 102; mental action, alternative account of 106–113; phenomenal 2, 114n1
constellations of powers 247
constitutive means 31
content condition 34, 37–39
content plurality in mental actions 31–39; complex contents 55; explanation of 40–46; inference 51–54; jointly sufficient conditions 39–40; judgments and decisions 49–51; objections to 54–55; overcrowding 54–55; philosophical importance of 46–54; possibility of 39–46; transparent self-knowledge 47–49
covert attention 71; auditory 71–72; intentional attunement works in 73; visual 71
covert perceptual attention 61, 71
critical reasoning, reasoning and 24
cross-cut non-agential behavioral characterization 212

Davidson, D. 19, 103, 179
daydreaming 88, 104
decision-making 244–245
decisions: judgments and 49–51; as momentary mental action of intention formation 256–259; neural 267–268; as nonactional 255–256; picking 261–264; proponent of views 264–265; studies and questions 259–261; theoretical and practical 49–50; *see also* practical decisions
deduction: attention, and mental action 148–153; capacity for 144;

epistemic content and phenomenal character of 152
degenerate selection 85–86, 96n12; aware of potential threat 86; selection for action 87; type of 86
Dehaene, S. 267–268
deliberation, structure of 143–148
deliberative version of Problem 63–64
Dennett, D. C. 23
derivative power: conjunctions of powers and constellations of powers 247; objective chance and 247–248; reasons explaining manifestations 234–237; reducing two-way powers 241–246; schema for 242; two-way powers as 228–249; unavoidability of substance dualism 238–241; *valence* of 243
desire 110, 111, 132–138; to act in accordance with reasons 15–16; addict and 15; in causal explanations of action 16; first-order 14; second-order 14–15
dichotic listening 71
Dietrich, F. 214, 284–286, 294
Dilthey, W. 198
directed acyclic graphs (DAGs) 172, 173, 175, 178
disappearing agent: Brent 18; moral of story 19; Velleman 14–16; Wallace 16–17
distinctively mathematical explanation 164; generation of 170–171; naive action explanation and 168–171
"dual method" hypothesis (DMH) 187
Duckworth, A. 295

economics, mental generalizations in 215
efficacy of action: causal Markov and faithfulness justifying inferences in 171–173; causal modeling and 164–180; model versus system as primary target of inquiry 168–171; naive action explanation and rationalization 166–168; Rationalization condition 174–179
effort *see* exerting effort; power of effort
Embodied Cognition 206; and causal roles of mental 205–208; evolutionary argument for Semantic Efficacy 218–222; Semantic Efficacy and Classical Cognition 208–211; Semantic Efficacy and special sciences 211–218
enkratic aversion of akrasia 277
entailment of attention 8; to determination 90–95; by mental action 87–90
episodic memory 109, 110
evaluative control 156
exerting effort 102, 109–111; alternative account of conscious mental action and 106; causal power and 19, 110; delivering content to consciousness by 111; disjunctive 110; relevant cognitive capacity and 105, 111
explanandum 166, 167, 190
explanation: content plurality in mental actions 40–46; involving mental action 5–6; *see also* causal explanation; distinctively mathematical explanation; naive action explanation
explanatory and metaphysical worries: reasons explaining manifestations of two-way powers 234–237; unavoidability of substance dualism 238–241
explanatory transparency 196–197

faithfulness 171–173, 219
feature-integration theory 81
Foran, S. T. 21–22
four-fold classification of lexical verbs 157n4
Frankfurt, H. G. 19
Frege, G. 52
functional role, attention 94

Gibbard, A. 50
Gibbardian expressivism 51
go-signal study 262–263
Godfrey of Fontaines 236, 237
Gould, C. G. 23
Gould, J. L. 23
Govier, T. 133–134

Hampshire, S. 129–130
Harman, G. 143
Hausman, D. M. 172
Heil, J. 123–124, 133

Henry of Ghent 237
heterogeneity of attention 80; attentional adverbialism 81–83; overreacting to 81–83; ubiquity and 80–87
Hieronymi, P. 155–156
Holton, R. 295
Holyoak, K. J. 24
homomorphism 208–209, 217
human agency: as biological phenomenon of philosophical significance 67; biology of 64; mental action and 26–28
human behavior: and action 25–26; modeling 176
Humeanism, defense of 273
Hunter, D. 5, 10, 122–138
hydraulic model 16–17
Hyman, J. 91, 123, 137

imperceptibility, property of 241
inference 145–146; achievements and 128; acting on oneself 129–132; actions, causings, and agents 123–126; causal Markov and faithfulness justifying in 171–173; choice, and desire 132–138; duration and 126–127; ideas about 122; as mental action 122–138, 153–157; nature of 51–54; non-voluntary 136; and onset of believing 126–129; para-mechanical idea of 128; as performance 128; rational voluntary 135; Taking Condition 52; theories of 51–54; voluntary 136
intellect, will and 236–237
intellectual attention 151
intellectualism 236
intelligence 198–201
intentional action 40, 69; attention by intention in 68; behavior space and 69; development of capacity for 77n6; emerging from convergence of problems 64; event-causalist accounts of 18; notion of 103; producing 18; reasoning as 159n11; standard accounts of 100; standard story of 14
intentional-action-first approach 69
intentional agency 228, 235, 239; agent's capacity for 229; concept of 229; expressing 110; intuitions about 246; two-way powers and 230, 233–234, 238
intentional content 70–71; attentional selectivity to 76; delivery, involuntary content delivery and 107; mental action and 61; movement of mind 75
internal mechanism-environment interactions 219
internal practical conflict: Brent's proposal 18–19; Velleman's proposal 15–16; Wallace's proposal 16–17
'internal' utterances of sentences 54–55
internalist explanatory strategy 209–210
interventionism 164, 165
intracranial explanations of cognition 213
intrinsically attentional deeds 82, 87, 88, 95n5
Inzlicht, M. 273, 281, 295

James, W. 61, 65, 76, 89, 272
jointly sufficient conditions: explanation of content plurality 40–46; on performing mental action with content plurality 39–40; types of mental action 39–40
judgments: attention preferences and 292–294; bridging gap between 52–53; complex contents in 55; and decisions 49–51; executing intentions in 52–53; form of inference 52; and intention 257; 'internal' utterances of sentences 54–55; theoretical and practical 49–50

Kable, J. W. 295
Kaufman, A. 255
Kennett, J. 279, 280, 282, 292–293
Kim, J. 179, 232
knowledge-first epistemology 160n21
Kornblith, H. 183–184
Krajbich, I. 286
Kurzban, R. 295

Lange, M. 164, 168
Levy, Y. 5, 8, 79–95, 100–101
Libet, B. 260, 262, 266–267
linguisticism, error of 240
List, C. 19–20, 214, 284–286, 294

listening 92
Loose, S. M. 286, 287
loss of control threat 274–276
Lotka-Volterra toy model 169–171, 176
Lowe, E. J. 233, 235

Macrae, C. N. 295
Malebranche, N. 89
Many-Many Problem (MMP) 62–65, 84–85; attention as emerging from solving 65; in degenerate scenarios 87, 95–96n10; expansive conception of solving 86; necessity of solving 85; non-deliberative 67; and paying attention 85; types of 77n4
Martin, C. B. 240
Mayr, E. 230, 233–235
meditation 105
Mele, A. R. 9, 10, 149–150, 229, 255–268, 292–293
memory: episodic 109, 110; person's name 150; searching 150; spatial 213; *see also* working memory
mental action 61; action capacities 68–71; alternative account of 106–113; as attending 71–76; attention 7–8, 65–68; ballistic process 102–106; with content plurality 32–39; contentful 37, 39; deduction, attention, and 148–153; defined 3–5; deliberation, structure of 143–148; executing several intentions at once 36–37; executing several intentions with content conditions 37–39; explanations involving 5–6, 19; implications for genuine for related issues 10; inference as 122–138, 153–157; intentional 32–35; kind of 32; Many-Many Problem 62–65; performed to satisfy more than one intention at once 35–36; power of effort 100–114; practical decisions as 255–268; as rational glue 26–28; reasoning and 142–157; reduced to non-agential mental states or events 6–7; science challenging, phenomenon of genuine 8–10; state of play for 2–3; steps narrowing 32–39; *see also* conscious mental action; content plurality in mental actions
mental agency 80, 88; attention, as form of 282; implication of attending in exercises of 89; interest in 94–95; role in 7; scope of 88, 95n1; settling question of 143; ubiquity in 87
mental ballistics model 150
mental, causal roles of: Embodied Cognition and 205–208; evolutionary argument for Semantic Efficacy 218–222; Semantic Efficacy and Classical Cognition 208–211; Semantic Efficacy and special sciences 211–218
mental causation *see* causation
mental content: attributing causal efficacy to 211; body and environment metaphysically relevant to 205–207; causal efficacy of 208; in causal explanations 212; causal inefficacy of 209; and cognition 206; in cognitive sciences 213–214; level of 215; predictive power of causal generalizations 209; relational facts for determining 221; scientific practices appeal to 207–208; structuring of 83
mental processes 2–3, 87, 193, 205–206; causal properties of 209; Classical Cognition and 206, 207–208; computational account of 210; content-involving descriptions of 216, 221; contentful characterizations of 213, 214; Embodied Cognition hypothesis 207; explanatory strategy of 209; higher order 186; intracranial computational processes 209; mental contents and content-sensitive 211; natural bases of 223; project of explaining 217; reasoning as 129; resistance to specifications of 210; semantic descriptions of 215, 217–218; subconscious 93
mental representation 207; contents of 210; importance of 210; internal processing of 221; intracranial features of 216
mental resources, assessment 296
mere behavior, action and 19
metaphysical dependence 206

metaphysical relevance, transitivity of 206–207
metaphysical worries, explanatory and: reasons explaining manifestations of two-way powers 234–237; unavoidability of substance dualism 238–241
metaphysics of agency 228
Metzinger, T. 26
Miller, J. 263–264
mind-wandering 104
mindfulness 105
Miracchi, L. 205–223
Mischel, W. 293–294
model-system, as primary target of inquiry 168–171
Mole, C. 80, 82, 83, 94
Molnar, G. 228, 242–243
Morgenbesser, S. 261
most general factive mental state 5, 80, 91, 94
most general mental act 79–80; attention as 87–95; attention by mental action, entailment of 87–90; attentional adverbialism 81–83; entailment to determination 90–95; explanation of attention as 80; heterogeneity, overreacting to 81–83; selection for action 83–87; ubiquity and heterogeneity of attention 80–87; ubiquity, overreacting to 83–87
motivational powers, life without 272, 294–296
multi-track one-way powers 231
Muraven, M. 295
Myers, J. 295

naive action explanation 178–179; causal explanation and 167–168; context of 177; distinctively mathematical explanations and 168–171; explanatory load in 166; kinds of 177; Lotka-Volterra toy model in 169–171; nature of 167; and rationalization 166–168; to strengthen causal modeling inferences 164
Navon, D. 296
Neumann, O. 65, 66, 84, 89
neural decision 267–268
neuroscience, philosophy and 212–213
Nickel, P. 136
Nisbett, R. 186–188, 190, 193–195
Noë, A. 205
non-agential systems 19
non-agential view, attention and 279–280
non-deliberative version of Problem 63–64
non-derivative powers 242

object of action, identification of 89, 90
objections to mental action 54–55; complex contents 55; overcrowding 54–55
one-way powers 231–232; multi-track 231; pleiotropic 231–232; polygenic 232; two-way powers and 232–233
*The Opacity of Mind* (Carruthers) 185
Orquin, J. L. 286, 287
O'Shaughnessy, B. 264–266
overcrowding 54–55
Owens, D. 132–133
ownership, objects of 131
owning, believing and 131

paradox of self-control 277–279
paramecium, physiological basis of taxes in 20–21
Peacocke, A. 4, 10, 31–56
Penn, D. C. 24
perceptual attention 80, 152; covert 61, 71; episodes of 89; importance of 286
performance, inference and 128
Pettit, P. 19–20
phenomenal consciousness 2, 114n1
"phylogenetic refinement" approach 221
physicalism: non-reductive 238, 251n22; substance 238
picking situation 261–264
planning agents 24–25
pleiotropic one-way powers 231–232
polygenic one-way powers 232
position effects 186
Povinelli, D. J. 24
power of effort: alternative account of 106–113; ballistic process 102–106; mental action and 100–114
practical conception 41–43
practical decisions 255–259; as momentary mental action of

intention formation 256–259; as nonactional 255–256; picking 261–264; proponent of views 264–265; studies and questions 259–261
practical knowledge 41
practical rationality 27–28; production of behavior in terms of 22; uncertainty and conflict and 26
predator-prey relationships 176
preference-action link: intentionally breaking 290–292; mediated by attention 285
preference determines action principle 277–279
preferences: action and attention 286–290; action link, intentionally breaking 290–292; and associated priority structures 283–286; attention and judgments 292–294; preference determines action principle 277–279
premise-belief 147–148
premotor theory of visual spatial attention 73–74, 77n8
primitive agency 20–21
priority orderings 282–283
priority structure: view 282; preferences and 283–286
Problem of Selection 76n2; intelligence 198–201; intelligibility 198–201; processualism about 188–191; transparency and 195–198
processualism: about self-understanding 188–191; skepticism and 191–195
psychic energy 272
psychological agents 25
psychological explanation 171
psychological research program 81, 82
pure reflex, defined 63

Rangel, A. 286
rational glue, mental action as 26–28
rational voluntary inference 135
Rationalization condition 165, 174–176; causal variables, independence of 175–176; failure of 178–179; key to using 175; naive action explanation and 166–168; refraining from using 175; turning left 176–178
re-prioritization: account of self-control 280–283; model 295
reasoning: agent 237; contrastive 237; and critical reasoning 24; deduction, attention, and mental action 148–153; deliberation, structure of 143–148; explaining manifestations of two-way powers 234–237; inferring as mental action 153–157; as intentional action 159n11; and mental action 142–157; as playing influencing role 235; practical 25, 27, 49–50, 111; propositional 76
reductionism 213
Reisman, K. 170
remembering capacity *see* memory
representations 240; computational 209, 218; controlled features rely on 65; internal 208, 216–217; internal architecture and 222; intracranial 209, 218; level of neural 70; possibility of a consistent propositional 26; psychological states and 24; *see also* mental representation
robotics 214
Ryle, G. 126–128

Schmeichel, B. J. 295
Schneider, S. 238
Schurger, A. 267–268
Scotus, J. D. 237
selection for action theory of attention 63, 65–68, 83–84
self-attribute beliefs and intentions 47–49
self-attribution, transparent 47–49, 55
self-control: agents with special motivational powers 277; capacity for 276, 297n7; characterization of 276; conflict resolution 275–276; defined 274–277; diachronic strategies 276; full-blooded 292–293; intra-psychic strategies 276; loss of control threat 274–276; non-agential view of 297n6; paradox of 277–279; re-prioritization account of 280–283; role of attention in 280–281; significance of attention for 272; situational strategies 276; strategy 276; synchronic strategies 276;

types of 276; *see also* synchronic self-control
self-determination 183
self-distraction 291
self-interpretation, activity of 184
self-interpretative faculty 187
self-knowledge 113–114; epistemology 48–49; transparent 31, 47–49
self-understanding: case for 184–187; skepticism about 182–201
Semantic Efficacy 205; absent compelling empirical reason to reject 211; animal behavior 212; biological systems 219–220; and Classical Cognition 208–211; economics 212, 215; evolutionary argument for 218–222; linguistics 212; psychology 212; and Semantic Externalism 211; and special sciences 211–218
Semantic Externalism 206, 207, 211
Shepherd, J. 4, 6, 14–28
Sitt, J. D. 267–268
skepticism: case for 184–187; intelligence 198–201; intelligibility 198–201; processualism and 191–195; about self-understanding 182–201; self-understanding, processualism about 188–191; transparency and self-understanding 195–198
Smith, M. 280, 292–293
special motivational powers 272, 294–296
Spirtes, P. 173
split-brain subject 185, 187, 193
Sripada, C. S. 292–293
standard accounts of action 100, 107; challenges 106; defenders of 101; defense of 114n2; intentional agency and 110; proponent of 101
Steward, H. 230, 234, 235, 248
Strawson, G. 102–104, 106, 108–109, 111, 156
Strevens, M. 216
substance dualism 229; non-physical properties to 239; property dualism and 239; for proponent of two-way powers 238–241; unavoidability of 238–241
'suppositional' reasoning 158n5
synchronic self-control 294; agents engaging in 292; intentional 274, 283; non-intentional account of 274; paradox and 278; re-prioritization account of 273; strategies 279
'System 1' reasoning 157n1

Taking Condition 145–147
target of inquiry, model-system as 168–171
temporal mismatch: attempt to defuse 85; between decision and self-attribution 49; strategy for handling objection from 95n10; untenable 85
theoretical predictability 232
thermoregulation 219–220
Thompson, M. 164, 165–168
Thomson, J. 123, 127
Titus, L. M. 1–11
transparency and self-understanding 195–198
transparent self-knowledge 47–49
Trevena, J. 263–264
truism 279
two-way powers: concept of 228; conjunctions of powers and constellations of powers 247; as derivative powers 228–234; non-causal spontaneous powers 233; objective chance and 247–248; one-way powers and 232–233; ontologically irreducible 228; reasons explaining manifestations of 234–237; reducing 241–246; unavoidability of substance dualism for proponent of 238–241

ubiquity of attention 80; and heterogeneity of attention 80–87; overreacting to 83–87; selection for action 83–87
Ullmann-Margalit, E. 261
unconscious inference 157n1
universal mind-reading hypothesis (UMRH) 187
unwilling addict 14–15

Valaris, M. 5, 10, 142–157
Velleman, D. 14–16
Vendler, Z. 145, 155
Vendlerian classification 155
visual covert attention 71
visual imagination 104–105

visual spatial attention, premotor theory of 73–74, 77n8
volitionalist model 17
Volterra principle 169–170
voluntarism 236

Wallace, R. J. 16–17
Watzl, S. 7–8, 83, 272–297
Weisberg, M. 170
White, A. 82, 87, 124, 128, 147–148
will and intellect 236–237
William of, A. 236–237
Williamson, T. 80, 91, 94
willpower, as irreducible mental faculty 272
Wilson, T. 186–188, 190, 193–195
Woodward, J. 172
Wooldridge, D. E. 23
working memory: limits on 54; storage in 214
Wu, W. 4, 8, 9, 61–76, 80, 84–86, 89, 94, 281, 285–286

Zhang, J. 173